

# Journal of Applied Psychology

738 v.

Edited by

Donald G. Paterson  
University of Minnesota

---

## Consulting Editors

George K. Bennett, *Psychological Corporation*  
Harold E. Burtt, *Ohio State University*  
Allen L. Edwards, *University of Washington*  
Clifford E. Jurgensen, *Minneapolis Gas Co.*  
Irving Lorge, *T. C. Columbia University*  
Quinn McNemar, *Stanford University*  
Alexander Mintz, *City College of New York*

James P. Porter, *Claverack, New York*  
Harold F. Rothe, *Fairbanks, Morse and Co.,  
Beloit, Wis.*  
Julian B. Rotter, *Ohio State University*  
Edward K. Strong, Jr., *Stanford University*  
Donald E. Super, *T. C. Columbia University*  
Morris S. Viteles, *University of Pennsylvania*  
Alfred C. Welch, *Knox-Reeves, Minneapolis*

---

## Volume 37, 1953



Published Bi-monthly by the American Psychological Association, Inc.  
Price and Lemon Sts., Lancaster, Pa.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the act of March 3, 1879

Acceptance for mailing the special rate of postage provided for in paragraph (d-2), Section 34.40,  
L. & R. of 1948, authorized October 10, 1947

Copyright, 1953, by The American Psychological Association, Inc.

## Contents of Volume 37

### Articles

Anderson, S. B. Prediction and Practice Tests at the College Level.....	256
Anderson, S. B. Estimating Grade Reliability.....	461
Anikeeff, A. M. Factors Affecting Student Evaluation of College Faculty Members.....	458
Ash, P. and Hobaugh, T. R. Some Primary Ratable Characteristics of Instructional Films.....	293
Baehr, M. E. A Simplified Procedure for the Measurement of Employee Attitudes.....	163
Bass, B. M., Klubeck, S. and Wurster, C. R. Factors Influencing Reliability and Validity of Leaderless Group Discussion Assessment.....	26
Bass, B. M. and Wurster, C. R. Effects of the Nature of the Problem on LGD Performance.....	96
Bass, B. M. and Wurster, C. R. Effects of Company Rank on LGD Performance of Oil Refinery Supervisors.....	100
Beaver, A. P. Kuder Interest Patterns of Student Nurses.....	370
Beaver, A. P. Personality Factors in Choice of Nursing.....	374
Belson, W. A. The Effect on Recall of Changing the Position of a Radio Advertisement.....	402
Bendig, A. W. The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale.....	38
Bernberg, R. E. Socio-Psychological Factors in Industrial Morale: II.....	249
Bills, M. A. and Taylor, J. G. Over and Under Achievement in a Sales School in Relation to Future Production.....	21
Bridge, L. and Morson, M. Item Validity of the Lee-Thorpe Occupational Interest Inventory.....	380
Brown, C. W. and Ghiselli, E. E. Prediction of Labor Turnover by Aptitude Tests.....	9
Brown, C. W. and Ghiselli, E. E. Per Cent Increase in Proficiency Resulting from Use of Selective Devices.....	341
Brown, C. W. and Ghiselli, E. E. The Prediction of Proficiency of Taxicab Drivers.....	437
Callis, R. The Efficiency of the Minnesota Teacher Attitude Inventory for Predicting Interpersonal Relations in the Classroom.....	82
Canfield, A. A. Administering Form BB of the Kuder Preference Record, Half Length.....	197
Canfield, A. A., Comrey, A. L. and Wilson, R. C. The Influence of Increased Positive g on Reaching Movements.....	230
Canter, R. R., Jr. A Rating-Scoring Method for Free-Response Data.....	455
Case, H. W. An Analysis of Engineering Entrance Examinations.....	42
Cattell, R. B. and Anderson, J. C. The Measurement of Personality and Behavior Disorders by the I. P. A. T. Music Preference Test.....	446
Chriswell, M. I. Validity of a Structural Dexterity Test.....	13
Clark, K. E. and Swanson, C. E. Attitudes Toward Public Low-Rent Housing, Before and After Construction.....	201
Cohen, J., Vanderplas, J. M. and White, W. J. Effect of Viewing Angle and Parallax upon Accuracy of Reading Quantitative Scales.....	482
Coleman, W. An Economical Test Battery for Predicting Freshman Engineering Course Grades.....	465
Comrey, A. L. Group Performance in a Manual Dexterity Task.....	207
Drake, L. E. Differential Sex Responses to Items of the MMPI.....	46
Dunnette, M. D. The Minnesota Engineering Analogies Test.....	170

Dunnette, M. D. and Maloney, P. W. Factorial Analysis of the Original and the Simplified Flesch Reading Ease Formulas.....	107
Edwards, A. L. The Relationship Between the Judged Desirability of a Trait and the Probability that the Trait will be Endorsed.....	90
Edwards, A. S. The Relation of Light Intensity to Accuracy of Depth Perception.....	300
England, G. W., Thomas, M. and Paterson, D. G. Reliability of the Original and the Simplified Flesch Reading Ease Formulas.....	111
Fattu, N. A. and Mech, E. V. The Effect of Set on Performance in a "Trouble Shooting" Situation.....	214
Fleishman, E. A. The Description of Supervisory Behavior.....	1
Fleishman, E. A. The Measurement of Leadership Attitudes in Industry.....	153
Fleishman, E. A. A Modified Administration Procedure for the O'Connor Finger Dexterity Test.....	191
Forbes, F. W. and Cottle, W. C. A New Method for Determining Readability of Standardized Tests.....	185
Friedman, N. The Quartile Difference Method of Item Selection.....	356
Garry, R. Individual Differences in Ability to Fake Vocational Interests.....	33
Ghiselli, E. E. and Barthol, R. P. The Validity of Personality Inventories in the Selection of Employees.....	18
Gilliland, A. R. and Newman, S. E. The Humm-Wadsworth Temperament Scale as an Indicator of the "Problem" Employee.....	176
Gough, H. G. The Construction of a Personality Scale to Predict Scholastic Achievement.....	361
Gray, J. S., Sustare, G. and Thompson, A. An Apparatus for Measuring Operational Hand Steadiness.....	57
Harris, S. and Smith, K. U. Dimensional Analysis of Motion: V. An Analytic Test of Psychomotor Ability.....	136
Hay, E. N. A Note on Small Samples.....	445
Hendrix, O. R. Predicting Success in Elementary Accounting.....	75
Herzberg, F. and Russell, D. The Effects of Experience and Change of Job Interest on the Kuder Preference Record.....	478
Hofstaetter, P. R. The Actuality Measure in the Study of Public Opinion.....	281
Holland, J. L., Krause, A. H., Nixon, M. E. and Trembath, M. F. The Classification of Occupations by Means of Kuder Interest Profiles: I. The Development of Interest Groups.....	263
Irvine, D. A Note on Ranking Method.....	53
Iscoe, I. and Lucier, O. A Comparison of the Revised Allport-Vernon Scale of Values (1951) and the Kuder Preference Record (Personal).....	195
Jenkins, J. J. Some Measured Characteristics of Air Force Weather Forecasters and Success in Forecasting.....	440
Jenkins, W. L. An Index of Selective Efficiency (S) for Evaluating a Selection Plan.....	78
Jennings, E. E. The Motivation Factor in Testing Supervisors.....	168
Johnsgard, K. W. Check-Reading as a Function of Pointer Symmetry and Uniform Alignment.....	407
Kashdan, L. Efficiency of Tests When Used to Select the Better of Two Workers.....	345
Kinzer, J. R. and Kinzer, L. G. Predicting Grades in Advanced College Mathematics.....	182
Kriedt, P. H. and Gadel, M. S. Prediction of Turnover Among Clerical Workers.....	338
Kunnath, J. G. and Kerr, W. A. Function Analysis of Thirty-Two American Corporate Boards.....	65
Lawshe, C. H. and Nagle, B. F. Productivity and Attitude Toward Supervisor.....	159

Layton, W. L. Predicting Success in Dental School.....	251
LeShan, L. L. and Brame, J. B. A Note on Techniques in the Investigation of Accident Prone Behavior.....	79
Lincoln, R. S. Visual Tracking: III. The Instrumental Dimension of Motion in Relation to Tracking Accuracy.....	489
Long, L. and Perry, J. D. Academic Achievement in Engineering Related to Selection Procedures and Interests.....	468
Longstaff, H. P. and Jurgensen, C. E. Fakability of the Jurgensen Classification Inventory.....	86
MacLean, A. G., Tait, A. T. and Catterall, C. D. The F Minus K Index on the MMPI.....	315
Mausner, B. Studies in Social Interaction: III. Effect of Variation in One Partner's Prestige on the Interaction of Observer Pairs.....	391
McGurk, F. C. J. Socio-Economic Status and Culturally-Weighted Test Scores of Negro Subjects.....	276
McIntyre, C. J. The Validity of the Mooney Problem Check List.....	270
Moore, J. E. and Ross, L. W. The Changing of Mental Test Norms in a Southern Industrial Plant.....	16
Morgan, W. J. and Morgan, A. B. Logical Reasoning: With and Without Training.....	399
Mueser, R. E. The Weather and Other Factors Influencing Employee Punctuality.....	329
Murray, J. E. An Evaluation of Two Experimental Charts as Navigational Aids to Jet Pilots.....	218
Navran, L. Validity of the Strong Vocational Interest Blank Nursing Key.....	31
Nuckols, R. C. A Note on Pre-Testing Public Opinion Questions.....	119
Nuckols, R. C. A Study of Respondent Forewarning in Public Opinion Polls.....	121
Oliver, J. E. A Punched Card Procedure for Use with Partial Pairing.....	129
Parker, J. W., Jr. Psychological and Personal History Data Related to Accident Records of Commercial Truck Drivers.....	317
Peters, H. C. The Prediction of Success and Failure in Elementary Foreign Language Courses.....	178
Prothro, E. T. Identification of American, British, and Lebanese Cigarettes.....	54
Prothro, E. T. Identification of Cola Beverages Overseas.....	494
Remmers, H. H. and Kirk, R. B. Scalability and Validity of the Socio-Economic Status Items of the Purdue Opinion Panel.....	384
Rock, M. L. Visual Performance as a Function of Low Photopic Brightness Levels.....	412
Ross, S. and Fletcher, J. L. Response Time as an Indicator of Color Deficiency.....	211
Ross, S., Ray, W. and Della Valle, L. Pointer Location and Accuracy of Dial Reading.....	131
Schneider, D. E. and Bayroff, A. G. The Relationship Between Rater Characteristics and Validity of Ratings.....	278
Schofield, W. A Study of Medical Students with the MMPI: III. Personality and Academic Success.....	47
Shaffer, R. H. and Kuder, G. F. Kuder Interest Patterns of Medical, Law, and Business School Alumni.....	367
Siegel, A. I. and Siegel, E. Flesch Readability Analysis of the Major Pre-Election Speeches of Eisenhower and Stevenson.....	105
Simpson, R. H. Rating Patterns for Maximizing Competition and Minimizing Number of Comparative Judgments Necessary for Each Rater.....	290
Smader, R. and Smith, K. U. Dimensional Analysis of Motion: VI. The Component Movements of Assembly Motions.....	308

Smith, F. J. and Kerr, W. A. Turnover Factors as Assessed by the Exit Interview.	352
Smith, P. C. The Curve of Output as a Criterion of Boredom.	69
Springer, D. Ratings of Candidates for Promotion by Co-Workers and Supervisors.	347
Stacey, C. L. and Goldberg, H. D. A Personality Study of Professional and Student Actors.	24
Stanley, J. C. Study of Values Profiles Adjusted for Sex and Variability Differences	472
Swanson, C. E. and Fox, H. G. Validity of Readability Formulas.	114
Taylor, E. K. and Schneider, D. E. A Biasing Factor in Essay Response Frequency	288
Tomlinson, H. and Preston, J. T. Development of a Short Test to Predict a Complex Aggregate Score.	260
Torrance, E. P. Methods of Conducting Critiques of Group Problem-Solving Performance.	394
Tydlaska, M. and Mengel, R. A Scale for Measuring Work Attitude for the MMPI	474
Tyler, F. T. and Michaelis, J. U. A Comparison of Manual and College Norms for the MMPI.	273
Uhlaner, J. E., Gordon, D. A., Woods, I. A. and Zeidner, J. The Relationship Between Scotopic Visual Acuity and Acuity at Photopic and Mesopic Brightness Levels.	223
Weitz, J. and Nuckols, R. C. A Validation Study of "How Supervise?"	7
White, W. J., Warrick, M. J. and Grether, W. F. Instrument Reading III: Check Reading of Instrument Groups.	302
Willerman, B. The Relation of Motivation and Skill to Active and Passive Participation in the Group.	387
Woods, W. A. Influence of Ink Color on Handwriting of Normal and Psychiatric Groups.	126
Zuckerman, J. V. A Note on "Interest Item Response Arrangement"	94

### Book Reviews

Argyris' An Introduction to Field Theory and Interaction Theory: H. J. Eysenck.	327
Arsenian's In Memoriam—Rudolf Pintner: Donald G. Paterson.	499
Barlow's Mental Prodigies: Lewis M. Terman.	325
Campbell's Practical Applications of Democratic Administration: Hugh M. Shafer.	148
Curran's Counseling in Catholic Life and Education: Robert J. Sherry.	245
Deese's The Psychology of Learning: O. Hobart Mowrer.	433
Division of Occupational Analysis, United States Employment Service's Dictionary of Occupational Titles, Second Edition: Alan M. Kershner.	241
Dooher and Marquis' The Development of Executive Talent: C. G. Browne.	149
Dunsmoor and Davis' How to Choose That College: John W. Gustad.	326
Fredriksen and Schrader's Adjustment to College: John W. Gustad.	60
Gray's Psychology in Industry: Clifford E. Jurgensen.	63
Guetzkow's Groups, Leadership and Men: Abraham S. Levine.	244
Heneman and Turnbull's Personnel Administration and Labor Relations: A Book of Readings, and Pigors and Myers' Readings in Personnel Administration: Albert S. Thompson.	323
Hirsh's The Measurement of Hearing: Miles A. Tinker.	148
IES' Lighting Handbook, Second Edition: Miles A. Tinker.	59
Judd's Color in Business, Science, and Industry: Forrest L. Dimmick.	150
Karn and Gilmer's Readings in Industrial and Business Psychology, and Blum's Readings in Experimental Industrial Psychology: Philip H. Kriedt.	499
Kelly and Fiske's The Prediction of Performance in Clinical Psychology: Stanley E. Jacobs.	61

Kephart's The Employment Interview in Industry: Harold E. Burt	239
Laird and Laird's Practical Sales Psychology: S. Rains Wallace, Jr.	324
Lauer's Learning to Drive Safely: Stanley E. Jacobs	242
Maier's Principles of Human Relations, Applications to Management: Wilton P. Chase	432
Miller and Form's Industrial Sociology; An Introduction to the Sociology of Work Relations: Glaister A. Elmer	59
Parker and Kleemeier's Human Relations in Supervision: William E. Kendall	62
Prasad's Fatigue and Efficiency in Textile Industry: Harold F. Rothe	242
Shostrom and Brammer's The Dynamics of the Counseling Process: John W. Gustad	243
Steiner's A Practical Guide for Troubled People: Harold Seashore	500
Ulrich, Booz and Lawrence's Management Behavior and Foreman Attitude: Theodore R. Lindbom	433
Walker and Guest's The Man on the Assembly Line: John M. Cook	324
Wechsler's The Range of Human Capacities: James J. Jenkins	240
Weinland and Goss' Personnel Interviewing: Clifford E. Jurgensen	434
Welch and Stone's How to Build a Merchandise Knowledge Test: Edwin E. Ghiselli	64
Wolfe, Buxton, Cofer, Gustad, MacLeod, and McKeachie's Improving Undergraduate Instruction in Psychology: Sidney L. Pressey	147
Zaleznik's Foreman Training in a Growing Enterprise: Theodore R. Lindbom	63

### Applied Psychology in Action

Bills, M. A. Our Expanding Responsibilities	142
Hadley, H. D. The Non-Directive Approach in Advertising Appeals	496
Lindbom, T. R. Evaluating Supervisory Training at the Job Performance Level	428
Vallance, T. R., Glickman, A. S., and Suci, G. J. Criterion Rationale for a Personnel Research Program	429
A New Management Tool for Top Executives	321
Background of an Industrial Psychologist	321
How's Your Empathy?	431
Job Supervision of Young Workers	236
News Item	146
Noise and Absenteeism	322
Personnel Psychology in a Steel Company	238
Reading: Stop Wasting Your Time	498

### Miscellaneous

New Books, Monographs, and Pamphlets	151, 247, 328, 435, 502
--------------------------------------	-------------------------

## The Description of Supervisory Behavior \*

Edwin A. Fleishman

*USAF Air Training Command, Human Resources Research Center,  
Lackland Air Force Base, Texas \*\**

Previous research in the area of leadership has to a large extent been concerned with postulated traits that leaders should possess, or with over-all evaluations of leadership. The leader's actual behavior has been largely ignored. More recent research has concluded that leadership is to a great extent situational, and that what is effective leadership in one situation may be ineffective in another. It therefore seems desirable to have available a method of describing leadership behavior which can be applied to many different situations. If this were possible then different leadership patterns could be related to criteria of effectiveness in a wide variety of group situations in which leaders function.

There have been some recent attempts to develop methods for the description of leadership behavior. This article is concerned with one such attempt which was carried out within the framework of the Leadership Studies at the Personnel Research Board of Ohio State University. The primary emphasis in this article will be to describe the development of a Supervisory Behavior Description questionnaire for use in an industrial situation.

### Developmental Background of the Instrument

*The Leader Behavior Description.* The Supervisory Behavior Description is based on the Leader Behavior Description Question-

\* This study was carried out while the writer was at the Personnel Research Board, Ohio State University, in cooperation with the International Harvester Company.

\*\* Perceptual and Motor Skills Research Laboratory. The opinions or conclusions contained in this report are those of the author. They are not to be construed as reflecting the views or indorsement of the Department of the Air Force.

naire originally developed by Hemphill and the staff at the Personnel Research Board (2). The questionnaire contained 150 items which described *how* people in leadership positions operate in their leadership role.<sup>1</sup> The respondent marked for each item, how frequently the leader did what each item described (e.g., always, often, occasionally, seldom, never).

A major problem in this endeavor was the classification of the items into meaningful categories of leader behavior. The 150 items were derived from over 1,800 original items which were written and then classified by "expert judges" into the following nine a priori "dimensions" of leadership behavior:

1. Integration,—acts which tend to increase cooperation among group members or decrease cooperation among them.
2. Communication,—acts which increase the understanding and knowledge about what is going on in the group.
3. Production emphasis,—acts which are oriented toward volume of work accomplished.
4. Representation,—acts which speak for the group in interaction with outside agencies.
5. Fraternalization,—acts which tend to make the leader a part of the group.
6. Organization,—acts which lead to differentiation of duties and which prescribe ways of doing things.

<sup>1</sup> An earlier approach at the Personnel Research Board developed modified job analyses procedures for investigating types of organizational activities engaged in by persons in high organizational positions. These methods have been summarized by Stogdill and Shartle (5) and by Shartle (4).

7. Evaluation,—acts which have to do with distribution of rewards (or punishment).

8. Initiation,—acts which lead to changes in group activities.

9. Domination,—acts which disregard the ideas or persons of members of the group.

An example of an item assigned to the Integration area was "He encourages group members to work as a team." An example of one assigned to the Domination area was "He insists that everything be done his way."

Subsequent administration of the form yielded adequate reliabilities for the nine dimension scores (.71 to .88) when groups filled it out as describing their own leader. Moreover, group members were consistent in how they described the *same* leader. However, the striking feature of repeated use of the questionnaire in various types of situations was the lack of independence of the dimensions. Most of the intercorrelations were between .50 and .80.

Item analysis also showed that an item assigned to one dimension by a priori methods might just as easily correlate more highly with scores on dimensions to which the item was *not* assigned. Some reorganization of the items, into relatively more independent categories of leader behavior, therefore, seemed necessary.

*Factor Analysis and Revision of the Leader Behavior Description.* In order to identify empirically the factor structure of the questionnaire, a factor analysis of the items was undertaken.<sup>2</sup> The questionnaire was administered to 300 Air Force crew members who described their airplane commanders. The Wherry-Gaylord Iterative Factor Analysis Procedure (6, 7) was utilized in the analysis of the items.<sup>3</sup> The factors extracted were rotated to orthogonality and then to simple structure. The analysis revealed two major factors present, together with two minor fac-

tors. The major factors were defined as "Consideration" and "Initiating Structure."

Items in the "Consideration" dimension were concerned with the extent to which the leader was considerate of his workers' feelings. It reflected the "human relations" aspects of group leadership.

Items in the "Initiating Structure" dimension reflected the extent to which the leader defined or facilitated group interactions toward *goal attainment*. He does this by planning, communicating, scheduling, criticizing, trying out new ideas, etc.

The minor factors were tentatively labeled "Production Emphasis" and "Social Sensitivity."

### Pre-Test on an Industrial Population

New keys were developed to score the questionnaire along these factor dimensions. Items with the highest loadings and purest factor structure were selected for each key. It was felt that scoring the questionnaire along these four dimensions would yield lower intercorrelations between the dimensions and would thus give measures of more independent aspects of the leader's behavior. A 136-item *Supervisory Behavior Description* questionnaire was administered to a pre-test sample of 100 International Harvester foremen at the Company's Central School in Chicago. These foremen, representing 17 different plants, used the questionnaire to describe the behavior of their own supervisors. The questionnaires were scored along the new factor dimensions derived from the Air Force sample. The purpose of this industrial pilot-study was to find out how applicable these new scales were to the industrial sample, and to determine what further revision might be necessary.

*Dimension Reliabilities and Intercorrelations.* Intercorrelations of the dimension scores showed that they still had substantial overlap with one another when applied to this industrial population. The intercorrelations were between .56 and .80, with corrected split-half reliabilities between .77 and .95. It seemed possible that the categories of leader behavior which were most independent in this industrial sample might be somewhat different

<sup>2</sup> This analysis was performed by B. J. Winer under a Human Resources Research Laboratories contract directed by Hemphill at the Personnel Research Board.

<sup>3</sup> This procedure does not require the item intercorrelations, but starts with item-sub-test correlations. That this procedure yields the same factors as Thurstone's Centroid Method has been empirically demonstrated (6).

Table 1

Items Selected for the Revised Form of the Supervisory Behavior Description<sup>1</sup>

	Orthogonal Factor Loading			Orthogonal Factor Loading	
	"Consideration"	"Initiating Structure"		"Consideration"	"Initiating Structure"
<b>"Consideration"</b> Revised Key			<b>"Consideration"</b> Revised Key		
He refuses to give in when people disagree with him.	-.68	.06	He criticizes a specific act rather than a particular individual.	.63	.14
*He does personal favors for the foremen under him.	.40	.06	He is willing to make changes.	.78	.09
He expresses appreciation when one of us does a good job.	.70	.19	He makes those under him feel at ease when talking with him.	.86	.17
He is easy to understand.	.70	.13	He is friendly and can be easily approached.	.82	-.02
*He demands more than we can do.	-.40	-.08	He puts suggestions that are made by foremen under him into operation.	.87	.11
*He helps his foremen with their personal problems.	.32	.05	He gets the approval of his foremen on important matters before going ahead.	.65	-.02
*He criticizes his foremen in front of others.	-.49	.03	<b>"Initiating Structure"</b> Revised Key		
He stands up for his foremen even though it makes him unpopular.	.54	.08	**He encourages overtime work.	.20	.40
He insists that everything be done his way.	-.52	-.01	*He tries out his new ideas.	-.10	.42
He sees that a foreman is rewarded for a job well done.	.70	.05	He rules with an iron hand.	-.20	.58
He rejects suggestions for changes.	-.62	-.06	He criticizes poor work.	-.18	.59
*He changes the duties of people under him without first talking it over with them.	-.69	.09	**He talks about how much should be done.	-.20	.60
He treats people under him without considering their feelings.	-.72	.41	*He encourages slow-working foremen to greater effort.	.17	.33
He tries to keep the foremen under him in good standing with those in higher authority.	.68	.17	He waits for his foremen to push new ideas before he does.	-.07	-.28
He resists changes in ways of doing things.	-.57	.19	He assigns people under him to particular tasks.	.00	.26
*He "rides" the foreman who makes a mistake.	-.61	.37	He asks for sacrifices from his foremen for the good of the entire department.	.00	.46
*He refuses to explain his actions.	-.72	.23	He insists that his foremen follow standard ways of doing things in every detail.	.25	.72
*He acts without consulting his foremen first.	-.73	.01	He sees to it that people under him are working up to their limits.	-.17	.87
**He stresses the importance of high morale among those under him.	.73	-.11	*He offers new approaches to problems.	.36	.72
He backs up his foremen in their actions.	.62	.16	He insists that he be informed on decisions made by foremen under him.	.13	.51
He is slow to accept new ideas.	-.66	-.06	He lets others do their work the way they think best.	-.17	-.33
He treats all his foremen as his equal.	.66	.28			

<sup>1</sup> Items not starred used the format: 1. always; 2. often; 3. occasionally; 4. seldom; 5. never. Items preceded by an asterisk (\*) used the format: 1. often; 2. fairly often; 3. occasionally; 4. once in awhile; 5. very seldom. Items preceded by a double asterisk (\*\*) used the format: 1. a great deal; 2. fairly much; 3. to some degree; 4. comparatively little; 5. not at all.

Table 1—*continued*

	Orthogonal Factor Loadings	
	"Consideration"	"Initiating Structure"
"Initiating Structure" Revised Key		
**He stresses being ahead of competing work groups.	.03	.34
**He "needles" foremen under him for greater effort.	-.17	.50
He decides in detail what shall be done and how it shall be done.	.37	.63
**He emphasizes meeting of deadlines.	.10	.68
*He asks foremen who have slow groups to get more out of their groups.	-.22	.40
**He emphasizes the quantity of work.	.17	.51

from those found most independent in the Air Force data.

*Item Analysis.* In order to clarify this problem and to revise the questionnaire for industrial use a statistical analysis was carried out at the item level. Two kinds of information were obtained concerning each of the 136 items in the Supervisory Behavior Description questionnaire. First, the distributions of responses among the five choices for each item were considered. Second, tetrachoric correlations of every item with each dimension total score were calculated to give indices of the internal consistency of the dimensions and to reveal the sources of overlap between the dimensions. Thus, coefficients were not only computed between an item and its own dimension total score, but with each of the other three dimension total scores to which it had not been assigned.

This analysis revealed that most of the items correlated highly with the dimension to which they were assigned. However, it was also evident that most of the items correlated highly with one or more dimensions to which they were not assigned.

Following the Wherry-Gaylord rationale (6, 7), the item-dimension correlations were considered factor loadings of the items on the four oblique (correlated) dimensions. In order to

compare the loadings with those obtained from the Air Force population, transformation to orthogonality was accomplished and it appeared, by inspection, that this transformation brought the loadings more in line with the original factors derived from the factor analysis. Item loadings increased on dimensions to which they were assigned and decreased on other dimensions. This seemed especially true for the two major factors (Consideration and Initiating Structure). Further preliminary rotations were then made with the primary objective of rotating the items originally in the two minor factors into more independent clusters. It appeared that this might not be possible, and in the light of the high correlations between these factors and the other two, their utility as separate dimensions was questioned for this population. Practically all the variation could be accounted for by the two major dimensions.

*The Revised Questionnaire.* Based on the item-dimension loadings derived from this industrial population, two revised scoring keys were developed,—one for "Consideration" and one for "Initiating Structure." Criteria for item inclusion were: (1) the item should have a high loading with the dimension in which it was to be included; (2) the item should have as close to zero loading as possible on the other factor; and (3) items which did not discriminate among supervisors (most respondents picking the same alternates) were rejected.

Twenty-eight items best meeting these criteria for "Consideration" and 20 items for "Initiating Structure" were selected. Table 1 presents the items finally selected for the revised form. The loadings given are those derived from this industrial population.

It can be seen that most of the items assigned to each key have high loadings with that dimension and insignificant loadings with the other. In addition, one more step was carried out. It was possible to select items for the "Initiating Structure" key so that some items had small *negative* loadings on "Consideration," and others had small *positive* loadings on "Consideration." It was hoped that the total effect of this would be to cancel out further the unwanted variance in the "In-

itiating Structure" key due to these cumulative small loadings on "Consideration."

The items in each key were, as before, randomly distributed through the questionnaire.

#### Administration of the Revised Form

This 48-item revised Supervisory Behavior Description was then administered to another comparable sample of 122 foremen in one of the International Harvester Company's plants. Again they were to describe their own supervisors. Assurances were again given that no one in the company would see their answers.

Table 2 presents some of the results.

From the results on this sample, it appeared that the scores on the two dimensions were now independent of each other.

Another index of the utility of the instrument is the agreement among different respondents who describe the *same* supervisor's behavior. The variation in scores can be divided into that between descriptions of *different* supervisors and that within descriptions of the *same* supervisor. This "within description" variation represents lack of agreement between respondents describing the same supervisor. The analysis of variance revealed significantly less variation among descriptions of the same supervisor than between descrip-

tions of different supervisors.<sup>4</sup> This appears to be further evidence of the objectivity of this questionnaire procedure.

The questionnaire was also administered to a sample of 394 *workers* who described the behavior of their own foreman. In this case the reliabilities of the scales were .98 and .78 and the correlation between them was  $-.33$ . It will be recalled that the *pre-test* sample consisted of people at the foreman level, so it might be expected that the correlation between dimensions would be somewhat higher in this sample of workers. This correlation is still considerably lower than had been obtained between dimensions with previous forms of the instrument. An analysis of variance again revealed significant agreement among workers describing the same foreman.<sup>5</sup>

It appeared that the two dimensions isolated were quite meaningful in this industrial situation. Apparently a supervisor could be high in Consideration without necessarily being high or low in the amount of planning, pushing for production, scheduling, or initiating behavior engaged in. At least the usual "halo effect" from scale to scale that occurs in most instruments in this area, seems for the most part to have been eliminated. The independence of the dimensions has special relevance when one considers the relationships of each of the two dimensions with some external criteria of group effectiveness.

The development of external criteria of group effectiveness was far beyond the scope of the present study. However, the Industrial Relations Department of the plant did have available the number of labor grievances filed for each of 23 departments during an eight-month period. These were reduced to grievances per worker for each department and correlated against the mean scores derived (from descriptions by foremen) for the general foreman in charge of each department. Although the *N* of 23 is pitifully small, and the records attenuated by many uncontrollable factors,

<sup>4</sup> Peters and Van Voorhis (3) suggest the conversion of *F* ratios to epsilon ( $\epsilon$ ), a statistic which indicates the *strength* of relationship. For these results  $\epsilon = .65 (P < .01)$  for Consideration and  $.47 (P < .05 > .01)$  for Initiating Structure.

<sup>5</sup>  $\epsilon = .72 (P < .01)$  for Consideration and  $.64 (P < .01)$  for Initiating Structure.

Table 2

Means, Standard Deviations, Range, Reliabilities, and Intercorrelations of the Dimension Scores of the Revised Supervisory Behavior Description (*N* = 122)

	Consideration	Initiating Structure
No. of Items	28	20
Mean	82.3	51.5
Standard Deviation	15.5	8.8
Range <sup>1</sup>	22 to 106	13 to 68
Reliability <sup>2</sup>	.92	.68
Intercorrelation	-.02	

<sup>1</sup> In this form, the alternatives for each item were weighted from zero to four. Thus, the highest possible score was 112 for Consideration and 80 for Initiating Structure.

<sup>2</sup> Split-half correlations corrected to full length of each dimension by the Spearman-Brown formula.

correlations of  $-.43$  with "Consideration" and  $.26$  with "Structure" were obtained. Only the first coefficient is statistically significant. The trend, however, was for the high grievance departments to be those with supervisors lower in consideration and higher in the amount of structuring in their leadership behavior. These results, of course, are purely suggestive. A more highly controlled criterion study of group effectiveness, and relationships to these dimensions is a program of future research.

The instruments have also been found useful in evaluating a leadership training program for foremen in the company and in studying relationships of leader behavior with certain factors in the social situation in which the foremen operate (1).

### Summary

This paper has described the development of one approach to the problem of describing leadership behavior in industry. A questionnaire, based on earlier work by Hemphill, was constructed. By means of this questionnaire the leadership behavior of supervisors could be objectively described. The questionnaire measures two relatively independent leadership dimensions found meaningful in the industrial situation,—“Consideration” and “Initiating Structure.”

There is no implication in the study as to the degree of each kind of behavior that is desirable or undesirable. Recognizing the

situational nature of leadership, the need for relating these scales to effectiveness of particular kinds of groups in well-controlled criterion studies is stressed. Moreover, the study reported here was confined to supervisors in one particular company.

The questionnaire at present is regarded only as a research instrument for the study of leadership behavior. More research applying the scales to other industrial situations needs to be done before they can be more confidently assessed.

Received May 5, 1952.

### References

1. Fleishman, E. A. *Leadership climate and supervisory behavior*. Personnel Research Board, Ohio State University, 1951.
2. Hemphill, J. K. *Leader behavior description*. Personnel Research Board, Ohio State University, 1950.
3. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
4. Shartle, C. L. Leadership and executive performance. *Personnel*, 1949, 25, 370-80.
5. Stogdill, R. M., and Shartle, C. L. Methods of determining patterns of leadership behavior in relation to organization structure and objectives. *J. appl. Psychol.*, 1948, 32, 286-91.
6. Wherry, R. J., Campbell, J. T., and Perloff, R. An empirical verification of the Wherry-Gaylord iterative factor analysis procedure. *Psychometrika*, 1951, 16, 67-74.
7. Wherry, R. J., and Gaylord, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, 8, 247-64.

## A Validation Study of "How Supervise?"

Joseph Weitz and Robert C. Nuckols

*Life Insurance Agency Management Association, Hartford, Conn.*

The problem of whether or not "How Supervise?" is an intelligence test has been discussed in several recent articles. Millard<sup>1</sup> has briefly discussed these studies and has presented additional data showing a relationship between this test and intelligence. We have been interested in determining whether or not "How Supervise?" is a test of supervisory ability and incidentally have some findings which may be relevant to its relations with intelligence.

### Procedure

A modification of "How Supervise?" was taken by 78 District Managers in one life insurance company. These districts are located throughout most of the southern and border states. The Managers supervise and direct the work of varying numbers of agents, ranging from 8 to 100.

By arrangement with the Psychological Corporation the test was modified by taking items from forms A and B and combining them into a test of 100 items. We used 20 items from the section Supervisory Practices, 32 items from Company Policy, and 48 items from Supervisor Opinions.

The kinds of changes made in the test were these: "Admitting to your workers when you make a wrong decision," was changed to "Admitting to your agents when you make a wrong decision." "Requiring supervisors to submit in writing their reasons for firing or penalizing any employee," was changed to "Requiring Managers to submit in writing their reasons for firing or penalizing any agent." These were the only changes made; that is, substituting "Manager" for "supervisor" and "agent" for "worker" or "employee."

The 100 items were put together into one test which we called the Manager's Inventory. This test was mailed to 83 managers and, as we mentioned earlier, 78 of them returned

completed questionnaires. Here the second difference occurred, that of a change in the testing conditions. The instructions for each section were the same as in the original test with the exception that "Manager" was substituted for "supervisor." However, it was truly self-administered with no time limit. The Managers signed the questionnaire. This permitted validation against certain criteria data for each district.

### The Criteria

Many different criteria were used. These included three production criteria: production of ordinary insurance, industrial insurance, and ordinary increase. (For those of us who do not know much about insurance terminology, it should suffice to say that these are three measures related to volume of sales.) We used as another criterion the number of men who terminated in each district during 1951. This figure was corrected for size of the district. We also used the four-year turnover, again corrected for district size, for the period 1947 through 1950. (This criterion has an odd-even year reliability of .77.) Another criterion was the persistency of the business sold, i.e., the average lapse ratio for each district. This might be thought of as the quality of the business.

In addition to the above criteria we had certain biographical data on each Manager. The only part of this information which we will discuss in the present paper is the highest school grade completed.

### Scoring of the Test

A number of different scores were obtained for each part and for the total test. For the total and for each of the parts we obtained the number of items right, the number wrong, the number right minus the number wrong, and the number of question marks. The correct answers were obtained by using the key origi-

<sup>1</sup> Millard, K. A. Is *How Supervise?* an intelligence test? *J. appl. Psychol.*, 1952, 36, 221-224.

Table 1

Correlation of Scores vs. Criteria Measures

	Ordinary Production Per Man	Lapse Ratio Per Man	Ordinary Increase Per Man	Industrial Increase Per Man	1951 Turnover	4 Year Turnover	Education Level
Number Right							
Part I	-.08	.19	-.03	-.04	-.11	-.18	.41
Part II	-.12	-.10	-.05	-.05	-.11	-.23	-.02
Part III	-.01	.12	.02	-.10	-.06	-.19	.33
Total	-.08	.09	-.01	-.10	-.11	-.26	.31
Number Wrong							
Part I	.12	.07	.09	.09	.07	.08	-.24
Part II	.26	.00	.23	.14	-.13	.14	-.02
Part III	.20	-.06	.16	.19	-.05	.00	-.27
Total	.28	-.02	.24	.21	-.08	.09	-.25
Right-Wrong							
Part I	-.12	.11	.06	.07	-.11	-.18	.41
Part II	-.20	-.06	-.15	-.10	-.01	-.22	-.01
Part III	-.10	.11	-.06	-.15	-.02	-.13	.34
Total	-.19	.07	-.12	-.17	-.02	-.23	.34
Number of ?							
Part I	.01	-.24	-.03	-.01	.07	.14	-.28
Part II	.09	.11	-.14	-.06	.23	.15	.04
Part III	-.15	-.10	-.16	-.04	.12	.22	-.18
Total	-.12	-.08	-.15	-.05	.17	.21	-.16

 $r = .22$  significant at 5% level. $r = .29$  significant at 1% level.

nally devised for each of the appropriate items in "How Supervise?"

### Results

The results are shown in Table 1. It can be seen that most of the correlations are below the five per cent level of significance with the exception of the scores vs. education where more of the correlations are above the five per cent level than could be expected by chance alone.

After finding no over-all significant relationship between the scores and the criteria, we did an item analysis on half of the cases. Using high and low district termination rate as the criteria we isolated those items, about twenty in all, which seemed to differentiate these two groups to some extent. In those items which differentiated the groups, the an-

swers predominately given by the low termination group were scored as correct. We now applied our new scoring key to the other half of the sample. It did not cross-validate; on the other half of the sample there was no relationship between the score obtained with the new key and termination rate.

### Conclusion

If the minor modifications of the test did not change "How Supervise?" materially, it would look as if this test is not valid in this situation for predicting agent turnover or production, both of which we feel should be related to supervisory ability. From our results the only thing this test seems to relate to is educational (intelligence?) achievement.

Received November 28, 1952.

Early publication.

## Prediction of Labor Turnover by Aptitude Tests

Clarence W. Brown and Edwin E. Ghiselli

*University of California, Berkeley, California*

With the popularization of the finding from World War I of a positive relationship between occupational level and intelligence test score, the notion developed that for each occupation there is an optimal level of intelligence. This belief led to a series of studies, particularly during the 1920's, which in general showed a curvilinear relationship to exist between scores on intelligence tests and labor turnover. Those individuals on a particular job who earn intelligence test scores at approximately the average of the group tested tend to remain on the job a longer time than those who earn scores at either extreme (e.g. 5).

Little attention has been given to the problem of the relationship between labor turnover and scores on types of tests other than those of intelligence. With tests of "specific" aptitudes the primary interest has been in discovering the correlations between test scores and some criterion measures of job proficiency or success in training. It is possible that the criterion of length of service on the job might have a curvilinear relationship with scores on specific aptitude tests as length of service has been shown to have with intelligence tests. If this were the case then there would be reason to question the notion that optimal intelligence test scores for various jobs are indicative of the "intellectual requirements" of the jobs. Intellectual factors other than those measured by intelligence tests would have to be considered. The study reported here was undertaken to investigate the nature of the relationship between scores on tests not ordinarily considered intelligence tests and labor turnover.

### Methods and Procedure

The subjects used in the present investigation were taxicab drivers. At the time they applied for work they were administered a number of tests as a part of the hiring procedure. To some extent the scores on these

tests were taken into account in the decision regarding employment. But other factors such as age, nature of previous experience, and scores on an interest questionnaire also entered into the hiring decision.

Those men who were ultimately hired were divided into two groups, those who stayed on the job for three months or more and those who left in less than three months. No differentiation was made between individuals who were separated for cause and those who left voluntarily. The number of enforced separations was very small, and resignations appeared in some cases not to be wholly voluntary but rather as a means for avoiding disciplinary action. The only individuals not included were those who were terminated because of illness, called to the armed services, or transferred to other jobs within the company.

All of the tests utilized were of the paper and pencil variety. The tests are listed in the accompanying tables. All three arithmetic tests involved computations but varied in the complexity of the problems presented. The Speed of Reactions tests involved making differential checking responses in accordance with pre-established rules to presentations of letter stimuli varying in spatial organization. In Test I the rules were given on each page and in Test II the rules had to be remembered. In the Dotting and Tapping tests, scores were based upon the speed with which dots were placed in small printed circles by means of a pencil. In the Dotting test, precision of movement was more of a factor than in the Tapping test because the circles were much smaller in size. The Judgment of Distance test required judgments about the relative distance of pictured objects based primarily on cues of perspective and interposition of objects. The Distance Discrimination test required judgments about the relative lengths of lines. In the Mechanical Principles test, problems involving knowledge of mechanical

functions and principles were presented. A more detailed description of these tests has been given elsewhere (2).

All men did not take all tests. In the present analysis the numbers of cases per test ranged from 218 to 441. Scores on each test were transmuted into normal standard scores on a nine-point scale following the procedure utilized in the Aviation Psychology Program of the Air Force (3). In standardizing the tests on this scale all applicants were utilized, whether they were hired or not. The distributions of scores of cases utilized in the subsequent analyses are given in Table 1.

### Results

For each score on the various tests the per cent of individuals leaving the job in less than three months is given in Table 2. These data are shown graphically in Figure 1. For three of the tests, Speed of Reactions II, Judgment of Distance, and Mechanical Principles, no relationship of any kind is apparent between test scores and turnover. For the remaining tests, curvilinear relationships occur and tend to be of the U-shaped kind found with intelligence tests, that is, individuals earning either high or low scores are more likely to quit the job sooner than those earning scores in the middle of the range.

The most consistent and striking relationship between test scores and turnover holds for

Table 1

Numbers of Hired Taxicab Drivers Earning Various Scores on the Several Tests

Test	Score					
	1 to 3	4	5	6	7	8 and 9
Complex Arithmetic	24	66	38	41	25	24
Intermediate Arithmetic	26	34	56	48	30	29
Simple Arithmetic	37	41	47	44	23	26
Speed of Reactions I	53	77	93	83	62	52
Speed of Reactions II	55	82	94	78	58	53
Dotting	80	89	88	84	53	47
Tapping	76	96	91	76	57	45
Judgment of Distance	62	47	67	87	34	35
Distance Discrimination	72	61	105	92	56	55
Mechanical Principles	56	51	87	52	44	42

Table 2

Per Cent of Taxicab Drivers Leaving Their Jobs in Less Than Three Months in Relation to Scores on Various Tests

Test	Score					
	1 to 3	4	5	6	7	8 and 9
Complex Arithmetic	42	30	37	17	20	33
Intermediate Arithmetic	62	35	14	29	40	50
Simple Arithmetic	30	39	23	27	30	31
Speed of Reactions I	40	31	33	27	35	42
Speed of Reactions II	35	34	33	33	33	36
Dotting	41	39	25	32	26	40
Tapping	38	38	30	29	28	44
Judgment of Distance	31	40	37	40	35	37
Distance Discrimination	32	41	29	32	34	44
Mechanical Principles	43	24	38	42	25	40

the Intermediate Arithmetic test. For Complex Arithmetic, Simple Arithmetic, Speed of Reactions I, Dotting, Tapping, and Distance Discrimination, scores and turnover seem to be correlated to about the same degree. No relationship is apparent between turnover and scores on Speed of Reactions II, Judgment of Distance, and Mechanical Principles.

Scores on all three of the arithmetic tests are related to turnover, as are scores on three of the four speeds tests (Speed of Reactions I, Dotting, and Tapping), and one of the two spatial tests (Distance Discrimination). It is therefore difficult to arrive at any generalization concerning the general factors in the tests which give the best forecast of turnover.

On the seven tests that are related to turnover, the optimal score varies between 5 and 6. It is 5 or very close to that value for Intermediate Arithmetic and Simple Arithmetic; about 6 for Complex Arithmetic, Speed of Reactions I, Dotting, and Tapping; and 5.5 for Distance Discrimination. Thus the optimal score on each of these tests is equivalent to or a little higher than the average score of this particular group of applicants.

### Discussion

From the findings of the present study, it is apparent that scores on some tests which in content are quite different from intelligence tests are related to labor turnover in the same

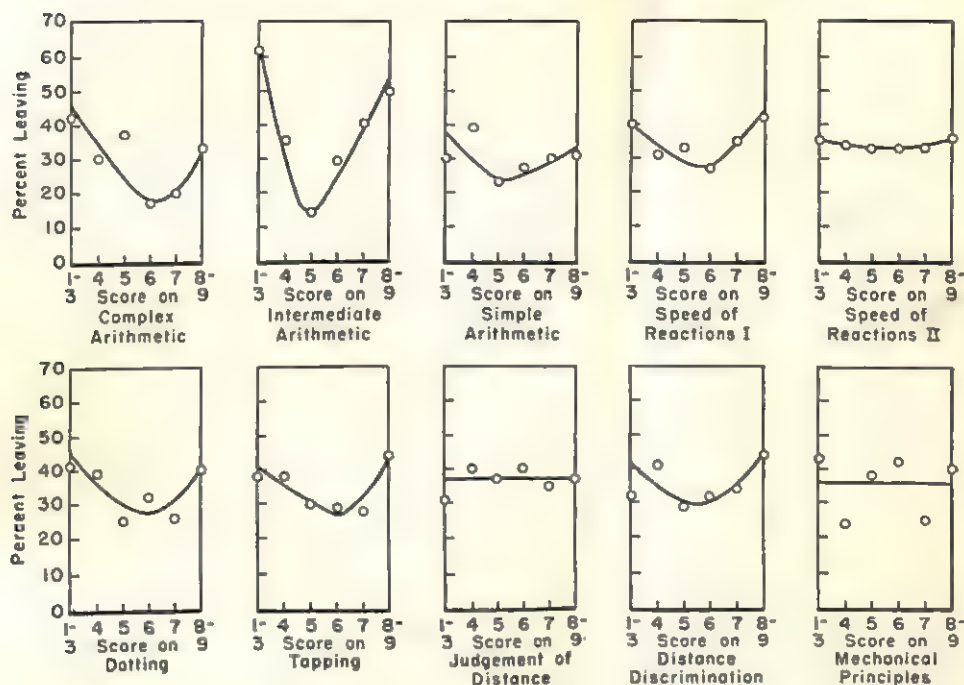


FIG. 1. Per cent of taxicab drivers leaving their jobs in less than three months in relation to scores on various tests.

manner as are scores on intelligence tests. Not only is the nature of the relationship the same, as a U-shaped relation, but the optimal scores, scores where turnover is at a minimum, fall at about the same place in the distribution of scores as do the optimal scores on intelligence tests. Some of the tests utilized in the present investigation, such as tapping and dotting, obviously measure quite simple functions which are unrelated to those measured by ordinary intelligence tests. The nature of the relationship with turnover, however, is the same.

The optimal scores on intelligence tests found in previous investigations, together with the relationship between intelligence test scores and occupational level, have been taken to indicate the "intellectual requirements" of jobs. In the present study we find the same kind of optimal scores with tests quite different from intelligence tests. Furthermore it has been found that even with tests of simple functions a similar relationship exists between scores and occupational level (4). It is not altogether certain, then, that the hierarchical levels of occupations with respect to intelli-

gence test score are to be accounted for solely on the basis of "intellectual requirements." Finally, it can be pointed out that in some instances turnover and intelligence test scores though correlated are not related in the U-shaped fashion. In Table 3 are data we have reported in a different form elsewhere concerning the relationship between intelligence test scores and turnover among bus drivers (1). In this occupation the large number of terminations was again the result of voluntary separation. From Table 3 it can be seen that turnover is at a minimum at the high

Table 3

Turnover Among Bus Drivers in Relationship to Intelligence Test Score

Score	N	% Leaving Under 6 Months
50-60	24	33
40-49	67	49
30-39	85	57
20-29	40	60
0-19	13	62

score levels and as the scores decrease there is an increasing rate of turnover.

It is not clear just exactly what generalizations can be drawn concerning the nature of the relationship between test scores and turnover. Certainly the use of the concept of "intellectual requirements" does not seem to be a satisfactory explanation. That is, the idea that turnover is a function of the distance, either plus or minus, of the person's intelligence from the mean intelligence for the job is not necessarily borne out. Just what types of aptitude tests give satisfactory predictions of turnover and what the nature of the relationship is between turnover and various aptitude qualifications is still obscure.

### Summary

Ten tests measuring several kinds of aptitudes were administered to groups of 218 to 441 taxicab drivers. For seven of the tests a U-shaped relationship was found between test scores and turnover, those individuals earning either high or low scores being more likely to

leave the job than those earning scores around the average of the group. Since this relationship is very similar to that found between scores on intelligence tests and turnover, it was concluded that the notion of "intellectual requirements" as an explanation of the U-shaped relationship between turnover and intelligence test scores is not wholly satisfactory.

Received May 2, 1952.

### References

1. Brown, C. W., and Ghiselli, E. E. Factors related to the proficiency of motor coach operators. *J. appl. Psychol.*, 1947, 31, 477-479.
2. Brown, C. W., and Ghiselli, E. E. Age of semi-skilled workers in relation to abilities and interests. *Personnel Psychol.*, 1949, 2, 497-511.
3. Flanagan, J. C. *The Aviation Psychology Program in the Army Air Forces*. Report No. 1, 1948, U. S. Gov't Print. Office.
4. Ghiselli, E. E. Intelligence test use in vocational guidance. In Kaplan, O. J. (Ed.) *Encyclopedia of Vocational Guidance*, Phil. Library, 1948.
5. Viteles, M. S. Selection of cashiers and predicting length of service. *J. Personnel Res.*, 1924, 2, 467-473.

## Validity of A Structural Dexterity Test

M. Irving Chriswell

*Buffalo Technical High School, Buffalo, N. Y.*

A test of Structural Dexterity emerged after a long period of testing in Technical High School of Buffalo, New York. Earlier experiments with the assembly of mechanical objects gave way to the construction of progressively complex structures of three dimensions.

In this test two different lengths of metal bars and pins comprised the unit parts. These were manually inserted and built upon a board divided into sections with holes drilled for each unit structure. The subject built each structure by interpreting the size and position of the parts from perspective sketches presented on a card. Features of the test follow:

1. A configuration of holes was adopted which became the basis for twelve different structures. The complete test utilized all areas of the board twice.

2. The progression from simple to complex structure gradually advanced the subject from one to two, and then from two to three level structures; from right angle to oblique positions; and from firmly built structures to movable balanced structures which required greater finger dexterity.

3. The score was determined by adding up the total number of pins and bars correctly placed. Deductions were made for errors. Testing time: 14 minutes. See accompanying photo in Figure 1.

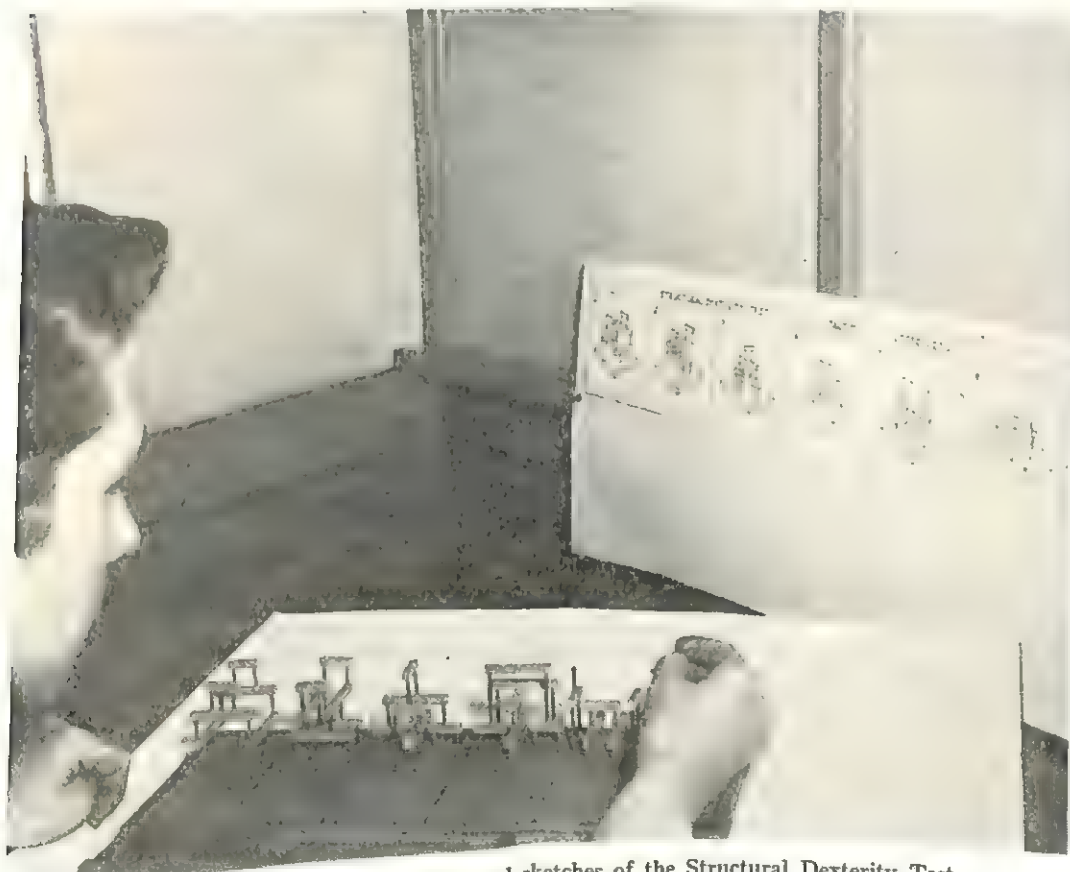


FIG. 1. Photograph of apparatus and sketches of the Structural Dexterity Test.

The Criteria

Five criteria were developed for this test. The first criterion, ( $C_1$ ), was the average of two ratings by a machine shop instructor. These ratings were not upon any specific job or project, but covered specific working traits, after three or four months of shop work. The second criterion, ( $C_2$ ), was the time in clock hours for the student to complete an assigned project of a small "C" clamp. Instructions were uniform and a detailed drawing and list of operations were furnished each student. The third criterion, ( $C_3$ ), was an averaged evaluation of layout, precision, and quality of work on the same "C" clamp by two judges, A and B, who were uninformed of the shop experience and behavior of individual students. The following scale provided the objective evaluation:

Clamp screw

- |                  |                |
|------------------|----------------|
| 1. Threading     | .....(0, 1, 2) |
| 2. Knurling      | .....(0, 1, 2) |
| 3. Total length  | .....(0, 1)    |
| 4. Knurl length  | .....(0, 1)    |
| 5. Hole, drilled | .....(0, 1)    |
| 6. Thread tested | .....(0, 1)    |
| 7. Chamfer       | .....(0, 1)    |

Total score .....

Clamp

- |                    |                |
|--------------------|----------------|
| 1. Outside contour | .....(0, 1, 2) |
| 2. Inside contour  | .....(0, 1, 2) |
| 3. Filing          | .....(0, 1, 2) |
| 4. Finish          | .....(0, 1, 2) |
| 5. Hole, true      | .....(0, 1)    |

Total score .....

Total, clamp and screw .....

Outside and inside contours were judged with the aid of a special steel template on a  $\frac{1}{16}$ " tolerance basis.

The fourth criterion, ( $C_4$ ), was the averaged evaluation,  $C_3$ , plus a time bonus. This bonus was developed from the time in hours for each job and was determined by the shop instructors: it weighted time compared with quality of work on a 1:2 ratio. The fifth criterion, ( $C_5$ ) was the shop teacher's evaluation on the objective scale plus the time bonus.

The Group and the Measures Used

A group was chosen which could be readily and precisely rated on their shop work. All pupils registered in first year Machine Shop were selected. This comprised a sub-group of 62 students in the 9th grade of the Electrical Course and another sub-group of 38 10th grade students of the Mechanical Course. Scores were available for these sub-groups on the following measures: Henmon-Nelson Intelligence Quotient, (IQ); Space and Numerical Ability, (DS and DN); on the Differential Aptitude Tests, (DAT); The Structural Dexterity Test described, (SD); the Purdue Pegboard, (PP), using the total score of Right plus Left plus Both Hands; and a test of Repetitive Operations, (RO), comprising nuts and bolts to be fastened to a block with the aid of an end wrench.

A comparison of the 9th and 10 grade sub-groups was undertaken and the means of the scores on the Structural Dexterity and the Purdue Pegboard tests were found to be significantly greater for the 10th grade than for the 9th grade sub-group. SD correlated with  $C_3$  . . . .44 for the 9th grade and .17 for the 10th grade sub-group. The Purdue Pegboard as well as both Differential Aptitude tests gave consistently low correlations (.08 to .30) with  $C_3$  for both sub-groups. A definite age difference of one year and three months existed between the sub-groups, and a marked difference in age correlations appeared: Age with  $C_3$  gave .15 for 9th grade and -.32 for 10th grade sub-group. Since these were the only unusual differences noted in the sub-groups, the combination seemed justifiable.

Reliability and Validity

Evidence of the reliability of the criteria was obtained. As previously stated, the third and fourth criteria were based upon the evaluations of two judges. The fifth criterion was based upon the evaluation of the shop instructor. Correlations of these evaluations follow:

	Judge B	Shop Instructor
Judge A	.78	.76
Judge B	—	.72

Table 1

Intercorrelations in the Prediction of Several Shop Success Criteria. Pearson formula used for all coefficients.  $N = 100$

	SD	RO	PP	DN	DS	IQ	Age	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>
SD		.48	.44	.13	.29	.16	.18	.38	-.38	.30	.41	.51
RO			.19	-.05	-.05	.12	.00	.25	-.43	.20	.35	.34
PP				-.02	.14	.14	.25	.10	-.31	.17	.26	.27
DN					.18	.48	-.05	.19	-.05	.04	.08	-.03
DS						.29	.11	.29	-.38	.23	.25	.21
IQ							-.37	.01	-.11	.00	.09	.08
Age								.08	-.13	.41	.30	.34
C <sub>1</sub>									-.33	.43	.49	.39
C <sub>2</sub>										-.38	-.78	-.69
C <sub>3</sub>											.82	.76
C <sub>4</sub>												.87
C <sub>5</sub>												

Using the correlation between Judges A and B, the Spearman-Brown formula yields a coefficient of .87 for the group of 100 students evaluated. This may be considered the reliability of the third criterion, and the minimum reliability of the fourth criterion.

The reliability of the SD test was determined by a method similar to the split-half technique. The coefficient for a group of 92 students in 9th and 10th grades was .88. Applying the Spearman-Brown formula the entire test would give .94.

An intercorrelation of factors and criteria is presented in Table 1.

With the exception of C<sub>2</sub>, the time criterion, it is significant that the Structural Dexterity Test has higher validity than any of the other selected tests. More significant results might be obtained with age held constant.

#### Summary

1. A test of structural dexterity shows significant differentiation in the performance of 9th and 10th grade technical high school students. The reliability by odd-even correlation employing the Spearman-Brown formula was .94 for a group of 92 students.

2. This test appears to be a valid measure of mechanical ability in a limited sense. It is a definite aid in the prediction of general machine shop success. The correlation for 100 subjects with averaged shop instructors' ratings on specific shop traits was .38; with time in hours to complete a specific job -.38; with averaged evaluation of a specific job by two judges .30; with this averaged evaluation plus a time bonus .41; and with a shop instructor's evaluation plus a time bonus .51.

3. This test of structural dexterity shows significant overlap with a test of repetitive operations (.48) and with the Purdue Pegboard, Right plus Left plus Both Hands score, (.44).

4. With multiple correlation formula, based upon the data presented, it was found that four selected factors, Structural Dexterity, Repetitive Operations, Space Relations (Differential Aptitude Battery) and Age, predicted the averaged evaluation plus the time bonus to the extent of .53. The multiple correlation between the same four factors and a shop instructor's evaluation plus the time bonus was .57.

Received March 24, 1952.

## The Changing of Mental Test Norms in a Southern Industrial Plant

Joseph E. Moore

*Georgia Institute of Technology*

and

Laurence W. Ross

*Union Bag and Paper Corporation, Savannah, Georgia*

In 1947 Bennett and Wesman (1) presented certain scores, which had been accumulated by Union Bag and Paper Corporation of Savannah, Georgia, on white men and women job applicants. The authors stated that for a given population local norms were the most meaningful. They also pointed out the often occurring problem of differences between local norms and "national" norms on which the test was originally based.

The problem of changes occurring in a given plant population from year to year naturally arises. It was our hope that the present study would throw some light on this subject. Does the level of performance, as measured by the Revised Beta Examination, remain relatively stationary for job applicants over a period of four or five years in a particular industrial plant?

The data on which this study was based cover a period from 1947 to 1951 and include all white men and women applicants who were given the Revised Beta Examination. The Union Bag and Paper Corporation requires all applicants who pass a preliminary interview to take a battery of psychological tests one of which is the Revised Beta Examination.

The subjects used in this study were 8,818 white men and 5,288 white women who applied for work at the Union Bag and Paper Corporation between the years 1946 and 1951. The average score (all scores used in this study are unweighted) earned by the men on the Beta Test was 83.8; the Standard Deviation for these scores was 15.9. The median score for the men was 84.5. The Stanford Binet Mental Age equivalent for this group of men applicants is 14 years (2). The Otis Self-Administering Test, Higher Examination,

Form A, score equivalent for the average of our group would be 33 points.

The average score for the women was 77.4 with a Standard Deviation of 15.8. The median for this group was 78.4. The Stanford Binet Mental Age equivalent for the women is 13 years, 1 month. A comparable score on the Otis Self-Administering Test, Higher Examination, Form A, would be 23.

The scores on the Revised Beta Examination for the groups in this study were compared with the scores obtained by Bennett and Wesman in an earlier study in the same plant in 1947 (1). Table 1 presents the reliability of the difference between the means of these groups of men applicants.

Table 1

Reliability of the Difference Between Mean Scores  
on The Revised Beta Examination for  
Men Industrial Applicants

Group	Num- ber	Mean	S.D.	"t"
Bennett & Wesman (1947)	1,362	80.5	17.7	6.32**
Moore & Ross (1951)	8,818	83.8	15.9	

\*\* Significant at .01 level of confidence.

In Table 1 it will be seen that the two groups of men industrial applicants are statistically significantly different. The mean mental test scores, however, earned by the 1951 men applicants is only 3.3 points higher than the mean of the 1947 group studied by Bennett and Wesman. This is less than one-fifth of the Bennett and Wesman S.D. of 17.7.

Table 2

A Comparison of The Mean Scores on the Revised Beta Examination of Two Groups of Women Applicants

Group	Num- ber	Mean	S.D.	"t"
Bennett & Wesman (1947)	1,083	72.9	17.5	7.75**
Moore & Ross (1951)	5,288	77.4	15.8	

\*\* Significant at the .01 level of confidence.

The 1951 mean is at the 55 percentile point on the 1947 percentile norms.

In Table 2 it will be seen that a difference in the mean scores of the women applicants on the Revised Beta Examination has also occurred. The 1951 group is performing at a higher level on the Beta Test than was true of the 1947 group. The 4.5 points difference in the mean is slightly larger for the women applicants than was found in the case of the men applicants. This 4.5 point difference is about one-fourth of the Bennett and Wesman S.D. of 17.5. The 1951 mean is also at the 55 percentile point on the 1947 percentile norms.

#### Summary and Conclusion

Bennett and Wesman reported on men and women applicants who took the Revised Beta

Examination prior to 1947. The data from these two investigators were compared with men and women applicants who took the examination between the years 1947 and 1951.

The present study shows that the men and women applicants seeking employment at this paper plant between 1947 and 1951 earned statistically significantly higher scores than did the group seeking employment prior to 1947. The difference between the mean scores of both the men applicants and the women applicants was, however, not striking, being about one-fourth of the 1947 S.D. The 1951 mean is at the 55 percentile point on the 1947 percentile norms.

The direction of the change is upward towards applicants who perform in such a way that they earn higher test scores on the Revised Beta Examination. The reason for these changes lies beyond the scope of this study.

Received April 14, 1952.

#### References

1. Bennett, George K., and Wesman, Alexander C. Industrial test norms for a southern plant population. *J. appl. Psychol.*, 1947, 31, 241-246.
2. Kellogg, C. E., and Morton, N. W. *Manual, revised beta examination*. New York: The Psychological Corporation.

## The Validity of Personality Inventories in the Selection of Employees

Edwin E. Ghiselli and Richard P. Barthol

*University of California, Berkeley, California*

Industrial and governmental organizations have for many years utilized tests of various kinds as aids in the selection of employees. Certain types of tests, e.g., aptitude, proficiency, and intelligence tests, have been shown to have merit in improving selection techniques. In recent years personnel workers have become increasingly conscious of the importance of personality factors as contributors to employee satisfaction or unrest. Personality tests and inventories have been used to supplement the subjective evaluation of these factors by the employment interviewers. A number of studies have been reported on the validity of personality inventories as selection devices, but these have been widely scattered through the literature. The purpose of this report is to summarize these studies so that the usefulness of the personality inventory can be more easily assessed.

### Methods and Procedure

In order to secure pertinent information we searched the various professional journals and books published from 1919 to date. From each study which reported findings concerning the validity of personality inventories for employment purposes, we noted the validity coefficient, the number of cases, and the job on which the group was employed. The studies included in the present analysis were restricted to those conducted in the United States, and to those in which the criterion was some index of job proficiency such as production records or ratings by superiors. An attempt was also made to include only those studies in which the scoring key for the personality inventory was developed independently of the group for which the validity coefficient was reported. Approximately 40% of the material reported in this paper is unpublished, having been drawn from various business, industrial, and governmental organizations.

In selecting the data for this study we examined the articles reporting the use of personality inventories and excluded those reporting traits that appeared to have little or no importance for the job in question. Thus, an inventory designed to measure sociability would be included for sales persons but not for machinists. We have grouped together all the remaining inventories regardless of the trait presumably measured. This was necessary because many utilize trait names that are very broad or not in general use, and some inventories do not name the trait at all.

### Results

In order to show the general trends the weighted mean validity coefficient was computed through Fisher's  $z$  for each of the major occupational groups. These values, together with the numbers of cases and numbers of validity coefficients, are given in Table 1. The distribution of the validity coefficients by occupation are shown in Figure 1.

There have apparently been few studies made on the efficacy of personality inventories for higher level supervisors. Contrary to expectations, the mean validity coefficient of only .14 is low and the distribution is some-

Table 1

Weighted Mean Validity Coefficients of Personality Inventories for Various Occupational Groups

Mean $r$	Total No. of Cases	Total No. of $r$ 's	Occupation
.14	518	8	General Supervisors
.18	6433	44	Foremen
.25	1069	22	Clerks
.36	1120	8	Sales Clerks
.36	927	12	Salesmen
.24	536	5	Protective Workers
.16	385	6	Service Workers
.29	511	8	Trades and Crafts

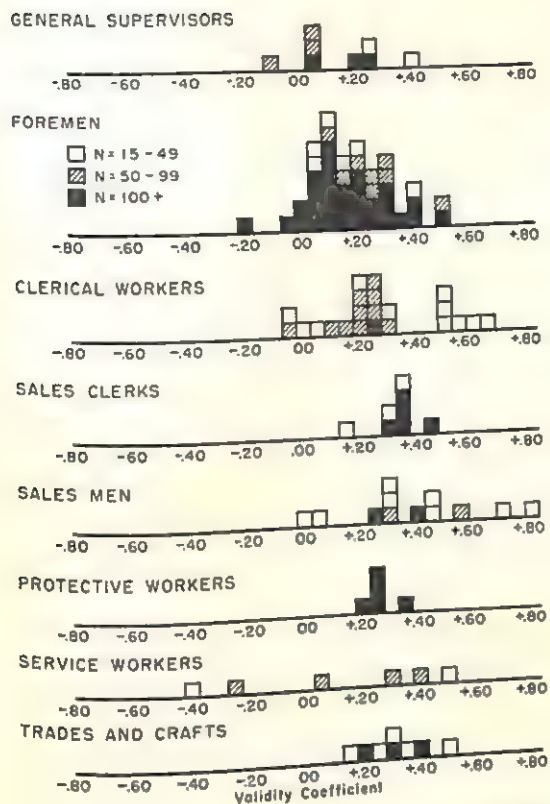


FIG. 1. Distribution of validity coefficients of personality inventories for various occupational groups.

what scattered. There is one case of fewer than 50 subjects with a substantial coefficient of correlation.

There were many studies reported for foremen which support the conclusion that personality inventories on the average do not have much predictive value in selecting supervisory employees. The mean and mode coincide at .18. Apparently certain inventories are used under certain conditions give good predictive results.

The studies made on clerical workers indicate that reasonably good predictions can be made on the basis of personality inventories. The mean value of .25 and the group of coefficients ranging from .50 to .65 demonstrate that this type of inventory can be seriously considered in devising a test battery for the selection of these workers.

For both of the sales groups, sales clerks and salesmen, quite substantial validities have been found. While there have not been as

many studies with these groups as might be expected the findings are fairly consistent. For both the mean validity coefficient is .36.

We found only five studies in which scores on personality inventories were related to proficiency among protective occupations. However, all of these studies utilized sizeable numbers of cases and are quite consistent in indicating moderate validity. The mean coefficient is .24.

In the studies of service workers the findings are quite inconsistent. Since the validity coefficients range from  $-.40$  to  $+.50$ , the low mean validity coefficient for this occupational group cannot be considered a representative indication of the effectiveness of personality inventories. It appears that under certain circumstances inventories may be used effectively.

The few applications of personality inventories to skilled workers have given quite promising results. The average of the validity coefficients for the trades and crafts is .29. Furthermore, the findings from different studies are quite consistent.

## Discussion

We were able to discover a total of 113 studies dealing with the validity of personality inventories in employee selection. When one recalls that these studies are spread over a number of different occupations it is apparent that the amount of information available for the evaluation of inventories is by no means extensive. However, a similar survey of reports concerning the validity of intelligence tests, certainly a much more popular instrument in employee selection, revealed only some 450 studies. Thus while in absolute terms the data may appear to be scanty, as compared to those available for other types of tests, they are fairly satisfactory.

It has been found that under certain circumstances scores on personality inventories correlate better with proficiency on a wider variety of jobs than might be expected. On the other hand there have been enough studies reporting negative results to emphasize caution in their use. These inventories have proved to be effective for some occupations in which personality factors would appear to be

of minimal importance (e.g., clerks, and trades and crafts), and ineffective for other occupations in which these factors could reasonably be expected to be of paramount importance (e.g., supervisors and foremen).

Received May 12, 1952.

### References

1. Beckman, R. O., and Levine, M. Selecting executives. *Personnel J.*, 1929-30, 8, 415-420.
2. Beckman, R. O. Ascendancy-submission test, revised. *Personnel J.*, 1932, 11, 387-392.
3. Diehl, H. S., and Paterson, D. G. *A personnel study of Duluth policemen*. Bull. Employ. Stabl. Res. Inst. Univ. of Minn., II, No. 2, 1933.
4. Dodge, A. F. Social dominance and sales personality. *J. appl. Psychol.*, 1938, 22, 132-139.
5. Dodge, A. F. What are the personality traits of successful clerical workers? *J. appl. Psychol.*, 1940, 24, 576-586.
6. Dornbusch, R. M., and Jones, M. H. *Handbook of employee selection*. McGraw-Hill, 1950.
7. Forlano, G., and Kirkpatrick, F. H. Intelligence and adjustment measurements in the selection of radio tube mounters. *J. appl. Psychol.*, 1945, 29, 257-261.
8. Freyd, M. Selection and promotion of salesmen. *J. person. Res.*, 1926, 5, 142-146.
9. Harrell, T. W. Testing cotton mill supervisors. *J. appl. Psychol.*, 1940, 24, 31-35.
10. Holmes, F. J. Validity of tests for insurance office personnel, I. *Personnel Psychol.*, 1950, 3, 57-69.
11. Holmes, F. J. Validity of tests for insurance office personnel, II. *Personnel Psychol.*, 1950, 3, 217-220.
12. Jurgensen, C. E. Report on the *Classification Inventory*, a personality test for industrial use. *J. appl. Psychol.*, 1944, 28, 445-460.
13. Kenagy, H. G., and Yoakum, C. S. *The selection and training of salesman*. McGraw-Hill, 1925.
14. Knauff, E. B. A selection battery for bake shop managers. *J. appl. Psychol.*, 1949, 33, 304-315.
15. Kurtz, A. K. Recent research in the selection of life insurance salesman. *J. appl. Psychol.*, 1941, 25, 11-17.
16. McMurry, R. M. Efficiency, work-satisfaction and neurotic tendency. *Personnel J.*, 1932, 11, 210-211.
17. Ream, M. J. *Ability to sell*. Williams and Wilkins, 1924.
18. Sartain, A. Q. Relation between scores on certain standard tests and supervisory success in an aircraft factory. *J. appl. Psychol.*, 1946, 30, 328-332.
19. Schultz, R. S. Standardized tests and statistical procedures in selection of life insurance sales personnel. *J. appl. Psychol.*, 1936, 20, 553-566.
20. Shultz, I. R., and Barnabas, B. Testing for leadership in industry. *Trans. Kan. Acad. Sci.*, 1945, 48, 160-164.
21. Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Co., 1940.

## Over and Under Achievement in a Sales School in Relation to Future Production

Marion A. Bills and Jean G. Taylor

*Ætna Life Affiliated Companies, Hartford, Conn.*

Beginning January 1, 1947, and based on previous experimental data, the Life Agency Department of the Ætna Life Affiliated Companies decided that it would require all applicants for selling positions to take three tests. These were: (1) Strong's *Vocational Interest Blank*; (2) the *Aptitude Index* published by the Life Insurance Agency Management Association (a scoring of an application blank and a personality test); and (3) *LOMA Test 1-A*, a mental alertness test published by the Life Office Management Association.

The Life Agency Department has regularly conducted schools for the training of agents. These schools have had various purposes and entrance requirements, but one series of schools conducted between January 1, 1947 and October 1, 1949, known as the "basic schools," were primarily for new agents and no previous selling experience or production was required for admittance. It has been noted from the first that there was a definite relationship between the *LOMA 1-A* test scores and the grades earned in the schools but that this relationship was by no means perfect. In addition, those who did better in the school than their test scores would indicate, seemed to be more successful in future selling. However, this result was not studied statistically until this year largely because the information came too late in our selection procedure to be of material benefit. However, since the information has at least borderline interest, might prove valuable in certain borderline cases, and since a large enough number of cases have been accumulated to justify a statistical study, we feel that it is advantageous to report certain of our results.

To keep the group as homogeneous as possible except for the two variables being studied, *LOMA 1-A* test scores and school grade, we limited the group to those who had scored an "A" on the Life Insurance scale of Strong's *Vocational Interest Blank* and an "A"

on the *Aptitude Index* and had attended a "basic school." There were 91 agents who met these requirements.

The grades in the "basic school" for these 91 agents ranged from 80 to 98 (S.D. = 3.96) with a mean of 90. *LOMA 1-A* test scores ranged from 99 to 209 (S.D. = 20.94) with a mean of 146. The correlation (product moment) between *LOMA 1-A* test scores and school grades was .64, between discrepancies and grade was .77, and between discrepancies and *LOMA 1-A* test scores was -.01. From the regression equation a predicted school grade was derived for each *LOMA 1-A* test score. This predicted score was then compared to the actual grade received to give an "index-of-achievement" score for each of the 91 individuals (actual grades minus predicted grades). This "index-of-achievement" score ranged from +7.8 to -6.6 and had a mean of 0 and a standard deviation of 3.03. Those receiving positive scores were considered "over" achievers, and those receiving negative scores "under" achievers.

The achievement scores were divided into three groups with extreme "over" achievers falling +3.0 and over, and extreme "under" achievers -3.0 and under. Table 1 gives the results of a comparison between the achievement scores and a combined criterion of length of service and premium production during the first year.

A Chi Square test for Table 1 yields a value of 15.44 which, with four degrees of freedom, is significant at the .01 level. It is evident that extreme "over" achievers, those with a score of +3.00 or over, in contrast to extreme "under" achievers, those with scores of -3.00 or less, tend more frequently to remain at least a year and to be higher producers. The one representative who was made an Agency Assistant before the end of the first year was an extreme "over" achiever and fell in the



Table 1

Over and Under Achievement in the Sales School *versus* Length of Service and Premium Production in the First Year

Index of Achievement (Actual School Grade Minus Predicted School Grade)	N	Per Cent of Agents Who Terminated Prior to 1 Year or Produced Less Than \$2500 or Both	Per Cent of Agents Who Remained 1 Year and Produced Between \$2500-\$4999	Per Cent of Agents Who Remained 1 Year and Produced \$5000 or Over
+3.00 and over	18	22%	39%	39%
+2.99 to -2.99	57	51	37	12
-3.00 and under	16	81	13	6

Table 2

Over and Under Achievement in the Sales School *versus* Length of Service and Total Two-Year Production

Index of Achievement (Actual School Grade Minus Predicted School Grade)	N	Per Cent of Agents Who Terminated Before the End of 2 Years or Produced Less Than \$5000	Per Cent of Agents Who Remained 2 Years and Produced \$5000-\$9999	Per Cent of Agents Who Remained 2 Years and Produced \$10,000 or Over or Became Agency Asst's*
+3.00 and over	18	22%	33%	45%
+2.99 to -2.99	57	55	30	16
-3.00 and under	16	75	19	6

\* Persons charged with agency management responsibilities and not engaged primarily in the sale of Life Insurance for their own accounts.

highest production group even with only eleven months of production represented.

In addition to success in the first year as treated in Table 1 we were also interested in following the same line of approach with total production during two years. Using the same breakdown of achievement scores but with a different division of the combined criterion of length of service and premium production, Table 2 was constructed.

A Chi Square based on Table 2 is 14.00 which, with four degrees of freedom, is significant at the .01 level. Table 2 indicates that the same general results persist over a two-year period.

In the above discussion, +3.00 and -3.00 were chosen as points where we could be reasonably sure that no chance variation in the school grading would account for the difference. However, it is of interest to note that

Table 3

Over and Under Achievement in the Sales School *versus* First Year Production

Relation of Actual School Grade to Predicted	N	First Year Production or Total Production if Left Under Twelve Months						No. and Per Cent of Agents Who Remained 12 Months or More (Includes 1 Made Agency Assistant)	
		Under \$2500		\$2500-\$4999		\$5000+			
		N	%	N	%	N	%		
		Over	Under						
	49	16	33%	19	39%	14	28%	43	88%
	42	29	69	12	29	1	2	26	62

Table 4

Over and Under Achievement in the Sales School *versus* Total Two Years' Production

Relation of Actual School Grade to Predicted	N	Total Two-Year Production or Total Production if Left Prior to End of Two Years						No. and Per Cent of Those Contracted Who Remained at Least Two Years (Includes Agency Assistants)		Made Agency Assistant (as of July, 1952)
		0-\$5000		\$5000-\$9999		\$10,000 and Over				
		N	%	N	%	N	%	N	%	
Over	49	17	35%	16	33%	16	32%	33	67%	11
Under	42	28	67	13	31	1	2	16	38	0

we obtain the same general results if the break between "over" and "under" achievers is made at zero although we find, as would be expected, that the differences are not as pronounced. These results are given in Tables 3 and 4.

#### Summary

The results reported in this study indicate that agents who receive a score of "A" on the Life Insurance Scale of Strong's *Vocational Interest Blank*, a score of "A" on the *Aptitude*

*Index*, and who achieve a higher grade in the "basic school" than would be predicted from their *LOMA* score: (1) remained with the company longer; (2) produced more paid premiums; and (3) were promoted to supervisory positions oftener than the agents who did not achieve a "basic school" grade as high as their *LOMA* test score would predict.

Received September 11, 1952.  
Early publication.

## A Personality Study of Professional and Student Actors \*

Chalmers L. Stacey

and

Herman D. Goldberg

*Syracuse University*

*Hofstra College*

For a long time it has been the contention of the authors that a great deal of frustration and heartbreak could be avoided if some measures of ability to attain success could be established for young people who wish to act in the professional field. Year after year thousands of young hopefuls flood the offices of the theatrical agents of Broadway and Hollywood determined to make a name for themselves. In many cases their decisions have been based on the fact that someone said that they were the possessors of pretty or handsome faces.

At the present time there is no standard for measuring or predicting the success a young, would-be actor hopes to attain. However the purpose of this experiment is not to establish such an over-all criterion but rather to determine some descriptive elements of the personalities of the groups studied.

Personality was selected as the basis of measurement for more than the reason of expediency. It was felt that personality as well as talent was one of the basic factors for attaining a certain amount of success in the field of acting. Discussion with the Broadway actors who were used as subjects in the study seemed to verify this fact.

It is hoped that the significant knowledge contained in the following material will be used to understand further the bewildering position of the young people who have chosen for themselves the difficult art of acting.

### Problem

The present experiment was designed in order to answer the following questions:

A. Do students in the School of Speech and Dramatic Art, Syracuse University, who express the desire to become professional actors and who appear in the major productions at

the Civic University Theatre, have a pattern of personality traits similar to that of professional actors?

B. Do students in the School of Speech and Dramatic Art, Syracuse University, who express the desire to become professional actors but for various reasons do not appear in the major productions at the Civic University Theatre have a pattern of personality traits similar to professional actors?

C. Do students in the School of Speech and Dramatic Art, Syracuse University, who express the desire to become professional actors and who appear in the major productions at the Civic University Theatre have a pattern of personality traits closer to that of professional actors than do students in the School of Speech and Dramatic Art, Syracuse University, who express the desire to become professional actors but for various reasons do not appear in the major productions at the Civic University Theatre?

### The Experimental Situation

*Subjects.* The following three groups were tested: (a) A total of 74 professional actors with a minimum of five years professional experience; (b) 30 students of the School of Speech and Dramatic Art who appeared in the University productions; and (c) 100 students of the School of Speech and Dramatic Art who did not appear in University productions.

*The Work of the Experimenter.* The experimenter presented the questionnaires to the subjects; insured no communication between subjects once the examination period began; and answered only questions concerning specific items.

*The Material.* Two personality questionnaires were used during the course of this experiment. These were J. P. Guilford's *An Inventory of Factors STDCR* which tested the factors: (S) Social Introversion-extraversion, (T) Thinking Introversion-extraversion, (D) Depression, (C) Cycloid Disposition,

\* The authors express their thanks for the assistance of Mr. Robert Breen of the American National Theatre and Academy; Mr. Clarence Derwent, President of Actors Equity; Professor Sawyer Falk, Director of Dramatic Activities at Syracuse University; and Mr. S. Eugene Perlman, assistant at Hofstra College.

Table 1

Showing the Variances for: (1) Professional Actors, (2) Student Actors Who Appeared in University Productions, and (3) Student Actors Who Did Not Appear in University Productions on Factors S T D C R and O Ag Co of the Inventories

	Group I	Group II	Group III
S	79.2	64.7	76.8
T	84.3	138.7	100.9
D	141.4	136.9	125.6
C	136.6	136.9	134.7
R	132.7	149.5	96.4
O	184.2	184.2	200.0
Ag	108.0	75.7	120.0
Co	190.2	226.7	345.5

Table 2

Showing "F" and "t" Values for the Differences of Variances and Means Among the Three Groups

	Groups I and II		Groups I and III		Groups II and III	
	"F"	"t"	"F"	"t"	"F"	"t"
S	1.1	2.8**	1.1	2.5*	1.2	.9
T	1.7*	3.2**	1.2	3.0**	1.4	1.1
D	1.0	2.6*	1.1	3.4**	1.1	.2
C	1.0	1.3	1.0	1.1	1.0	.5
R	1.0	1.6	1.4	4.5**	1.6	1.5
O	1.0	1.6	1.1	.7	1.1	1.1
Ag	1.4	.4	1.1	.5	1.6	.8
Co	1.2	.5	1.8**	1.3	1.5	.4

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

(R) Rhathymia; and The Guilford-Martin Personnel Inventory which tested the factors: (O) Objectivity, (Ag) Agreeableness, (Co) Cooperativeness.

*The Data and Their Analyses.* For the purposes of this study two statistical tests were used to analyze the data: the F test to test the differences in variances and the t test to test the differences in means. Tables 1 and 2 present these findings.

### Conclusions

The following conclusions were arrived at after an analysis of the data:  
On the factors, Cycloid Disposition, Objec-

tivity, Agreeableness, Cooperativeness, there was no difference in degree of personality trait between the professional actors and student actors who did not appear in university productions.

Professional actors, however, are significantly more shy, seclusive, and have a greater tendency to withdraw from social contacts than do student actors who do not appear in university productions (Factor S).

It may also be said that student actors who do not appear in university productions have significantly more of the tendencies to seek social contacts and enjoy the company of others (Factor S).

Professional actors are significantly more inclined to meditative thinking, philosophizing, analyzing one's self and others than are student actors who do not appear in university productions. These student actors tend significantly toward an extravertive orientation of the thinking processes (Factor T).

Professional actors show significantly more signs of depression than do students who do not appear in university productions. Furthermore, student actors who do not appear in university productions are significantly more cheerful and optimistic than professional actors (Factor D).

Professional actors are significantly more inhibited, over-controlled, conscientious and serious-minded than student actors who do not appear in university productions (Factor R).

Student actors who do appear in university productions and student actors who do not exhibit about the same degree of personality in all eight traits measured.

Professional actors and student actors who appear in university productions exhibit about the same degree of personality in traits of Cycloid Disposition, Rhathymia, Objectivity, Agreeableness and Cooperativeness.

The student actors who did appear in university productions were like professional actors on six and significantly different on only two of the eight traits measured. However, student actors who did not appear in university productions were like professional actors on four and significantly different on four of the eight traits measured.

Received March 10, 1952.

## Factors Influencing Reliability and Validity of Leaderless Group Discussion Assessment<sup>1</sup>

Bernard M. Bass, Stanley Klubeck and Cecil R. Wurster

*Louisiana State University*

A substantial body of evidence is available to suggest that behavior in the initially leaderless group discussion is indicative of leadership potential for a fairly wide range of situations (1, 2, 4, 5, 6, 8, 10, 11, 12, 13, 14). However, the leaderless discussion technique has one obvious handicap, at least. Since very few candidates can be placed in a single discussion, many discussions are likely to contain unrepresentative samples of the population being assessed. Some discussions may contain only persons high in leadership potential; others may contain only persons low in such potential.

Since it has been shown that a person's leaderless discussion rating will be lower, the more candidates he must compete with in a given discussion (3), it seems reasonable to hypothesize that a candidate's ratings will depend also to some extent on the "quality" as well as the quantity of those he is facing. The leadership potential and the effectiveness of discussion behavior among his fellow candidates will probably affect a candidate's rating even where raters attempt to use one standard for rating many discussions.

If these two sources of assessment error—variations from discussion to discussion in quantity and "quality"—were unrelated to variations in the reliability and validity of the LGD, they would cease to be a problem. On the other hand, if variations in the available leadership potential, in number of participants, and in effectiveness of discussion behavior of participants as a whole, were systematically related to the reliability and validity of the technique, then it would be profitable to isolate these errors and make appropriate allowances for them in the future. Also, the more reliability and validity of the technique were found to vary from discussion

to discussion, the more would these allowances be necessary. The purpose of this investigation was to determine the extent to which the reliability and validity of the LGD varied from discussion to discussion and the extent to which these variations could be accounted for by other known LGD variables such as the quantity and "quality" of participants.

### Method

The investigators had available the assessment and criterion data from LGD validation studies by Wurster and Bass (14) based on 14 discussions among fraternity pledges; by Doll (6) based on 20 discussions among sorority members and by Bass and Coates (2) based on 21 discussions among Army cadets and 12 discussions among Air Force cadets.

For each of the 67 discussions the following indices were computed:  $r_{cd}$  = the validity of LGD observers' ratings as indicated by their correlation with a criterion of peers' or superiors' appraisals of the participants' leadership potential;  $r_{xx}$  = the reliability of the two LGD observers' ratings as estimated by the correlation between them;  $M_d$  and  $SD_d$  = the mean and standard deviation of assigned LGD ratings;  $M_c$  and  $SD_c$  = the mean and standard deviation of participants' criterion ratings;  $K$  = the number of participants in a designated discussion; and  $E$  = the extent to which a designated discussion attained its objectives as rated by both observers on a five-point scale. The corrected split-half reliabilities of both observers' ratings of this measure of group effectiveness for the 4 studies respectively were .74, .75, .78 and .85.

<sup>1</sup> This study was aided by a grant from the Louisiana State University Council on Research.

<sup>2</sup> For a discussion of how these ratings were made the reader is referred to the original studies (2, 6, 14).

Table 1

Number of Groups, Means and Standard Deviations of  $r_{cd}$ ,  $r_{xx'}$ ,  $M_d$ ,  $SD_d$ ,  $M_o$ ,  $SD_o$ ,  $E$ , and  $K$ 

Subjects	No. Groups	Means							
		$r_{cd}$	$r_{xx'}$	$M_d$	$SD_d$	$M_o$	$SD_o$	$E$	$K$
Fraternity pledges (14)	14	.38	.85	39.7	14.9	40.7	15.9	5.6	6.2
Sorority members (6)	20	.17	.85	29.0 <sup>a</sup>	11.0 <sup>a</sup>	30.2	8.7	5.6	7.0
Army cadets (2)	21	.20	.81	41.2	17.1	19.2	5.6	5.4	7.1
Air Force cadets (2)	12	.42	.88	39.8	17.2	39.3	15.7	5.4	6.8
Weighted Average*		.27	.84	—	—	—	—	5.5	6.8
Standard Deviations									
Fraternity pledges (14)	14	.43	.09	7.0	3.8	6.9	3.0	1.6	.67
Sorority members (6)	20	.46	.10	3.4 <sup>a</sup>	2.6 <sup>a</sup>	4.3	2.5	1.7	.00
Army cadets (2)	21	.44	.15	4.8	3.1	2.5	1.7	1.6	.83
Air Force cadets (2)	12	.39	.12	6.3	3.6	11.4	6.0	1.9	.83
Weighted Average*		.43	.12	—	—	—	—	1.7	.58

\* Only computed for variables whose scales remained the same for all four samples of subjects.

<sup>a</sup> Used 7-item rather than 9-item rating scale.

### Variations in Reliability and Validity

Table 1 shows the means and standard deviations of the above measures study-by-study. Of special interest to the investigators were the following conclusions inferred from the results reported in Table 1.

1. The reliability of LGD observers' ratings appeared quite stable as judged from the standard deviation of these reliabilities of .12. Judging from the average mean reliability of .84, it appears that this measure suffered little when obtained discussion-by-discussion in comparison to previous studies where it was obtained from pooled data.

2. The average validity of the LGD of .27 was less than obtained for the same data when analysis was performed in previous studies by pooling many discussions. This suggested that the validity of the LGD would be lowered when use was made of any rating technique such as comparison rankings among discussion participants that did not make provision for between-discussion variations in the candidates.

3. The average standard deviation of the validity coefficients of .43 indicated that there was tremendous variability in the validities from discussion to discussion—almost as much variation as the correlation scale of +1.00 to -1.00 would allow. Therefore, it was deduced that any factors which could be found

to correlate with the validity coefficient, even if the correlations were quite low, could account for a wide variation in validity.

### Correlational Analysis

Pearson product-moment intercorrelations were computed between the various group measures. Each intercorrelation was computed separately for each of the 4 samples because of the variations in rating scales used from study to study. The correlations were transformed by means of Fisher's Z conversion and then averaged.

Several comments concerning this matrix shown in Table 2 appear pertinent:

1. The significantly negative correlation of  $-.37$  between group size and mean discussion ratings assigned corroborated similar results obtained by Bass and Norton (3), who varied group size systematically from 2 to 12. Thus, even where size differences were small and accidental as in the present analysis, substantial variations in LGD leadership ratings assigned were found associated with variations in the number of participants per discussion.

2. The significant correlation of  $.46$  between rated discussion effectiveness and mean LGD ratings assigned group-by-group suggested that the discussion observers had a consistent frame of reference in making these two ratings which transferred from one group to

another, since, by definition, leadership ratings of individuals were supposed to depend on the degree to which they moved their group towards its goal while group effectiveness was defined as degree of goal attainment.

3. Possibly the most valuable finding of this analysis was the significant correlation of .35 between the mean discussion rating assigned and the mean criterion status of discussion participants. This suggested strongly that absolute ratings of the discussion observers were accurately sensitive to group variations in outside leadership potential. It implied that there was a "between groups" positive correlation as well as a "within groups" positive correlation between LGD ratings and outside appraisals of leadership potential. It was the validity due to "between groups" covariance which was lost when correlational analyses between test and criterion were run group-by-group rather than for an entire sample. This probably accounted for the low mean validity of .27 based on group-by-group analyses reported in Table 1 in contrast to the validities of .40 and .50 reported when data are pooled.

4. The positive but insignificant correlation of .20 between the group-by-group standard deviations of discussion ratings and the group-by-group standard deviations of criterion ratings suggested that the observer's ratings were also somewhat sensitive to the variation in restrictions in range of the outside leadership potential among the participants. Once again, it is obvious that ratings which depend solely on standards within a group are most likely to suffer in validity. Rating techniques with this disadvantage include the forced distribution method, the paired comparison technique or the rank order-of-merit procedure where quotas, pairings or rankings are made within a designated group situation. Similarly, any ratings of each other by the candidates themselves will most likely be attenuated in validity since they will depend solely upon standards based on observation of a single discussion.

5. Despite the above relationships, while the means and standard deviations of criterion ratings were significantly positively related ( $r = .42$ ), the means and standard deviations of discussion ratings were significantly nega-

tively related ( $r = -.28$ ). This suggested that average and poor discussion participants were handicapped most severely when in competition with those participants of the entire sample who earned extremely high LGD ratings.

6. The reliability or extent of agreement between the two discussion observers appeared significantly related ( $r = .54$ ) with the standard deviation of the discussion ratings. However, this correlation was an artifactual relationship, since by a simple transposition of the formula for the standard deviation of the sum of correlated scores it can be shown that

$$r_{xx'} = \frac{SD_d^2 - SD_x^2 - SD_{x'}^2}{2SD_xSD_{x'}} \text{ where } x \text{ and } x' \text{ refer to each of the 2 observers' ratings and where } x + x' = d.$$

7. A significantly negative correlation of  $-.32$  was found between mean discussion ratings assigned and the reliability of ratings. A possible explanation for this correlation offered by the observers was that they, the observers, became more interested and absorbed in discussions with very good participants while remaining more detached when participants were poorer. This same hypothesis was used to account for the one highly significant curvilinear relationship which was found to exist among the variables. For each of the four samples respectively, etas of .54, .78, .73 and .68 were found between the rated effectiveness of the group discussion and the extent to which the observers agreed on the leader-

Table 2

Mean Intercorrelations Among  $r_{ed}$ ,  $r_{xx'}$ ,  $M_d$ ,  $SD_d$ ,  $M_o$ ,  $SD_o$ ,  $E$ , and  $K$  ( $N = 67$  Discussions)\*

	$r_{ed}$	$r_{xx'}$	$M_d$	$SD_d$	$M_o$	$SD_o$	$E$	$K$
$r_{ed}$								
$r_{xx'}$	.07							
$M_d$		-.32						
$SD_d$			-.28					
$M_o$				.06				
$SD_o$					.42			
$E$						.14		
$K$							.01	

\* With 65 d.f.,  $p < .05$  when  $r = .24$ ;  $p < .01$  when  $r = .31$ . All correlations significant at and below the 5 per cent level of confidence are in boldface type.

ship ratings they assigned. Agreement among observers reached a maximum in groups of average effectiveness or goal attainment while reliability of discussion ratings was low in both extremely effective and extremely ineffective groups.

8. The validity of the LGD was correlated at the 5 per cent level of confidence with three variables, the group-by-group means of criterion ratings, as well as the group-by-group standard deviations of discussion and criterion ratings. Since these three variables were all related to each other, it was difficult at this point to determine which ones were uniquely related with LGD validity.

### Multiple Correlational Analyses

It was decided to isolate the unique contribution to the variance of the validity of each of the other 7 variables of this investigation. This was done by determining the multiple correlation between the validity of the LGD and an optimally weighted sum of scores derived from the other 7 variables. The Doolittle rather than the Wherry-Doolittle solution was used to obtain the multiple R and the beta weights since interest of the investigators was focused on studying the effects of all the other variables on validity rather than the effects of the smallest number of the other variables which would yield the highest multiple correlation with LGD validity. The multiple correlation obtained was .43 indicating that approximately 19 per cent of the variance in LGD validity from group-to-group could be accounted for by those other variables. Of this 19 per cent, 6 per cent was accounted for by the standard deviation of criterion ratings; 5 per cent, by the standard deviation of LGD ratings; 4 per cent, by mean criterion scores; 3 per cent, by group size; and 1 per cent, by group effectiveness. These results suggest that:

1. The validity of a given discussion clearly suffered when there was a restriction in range of criterion ratings. This particular handicap will probably always remain with any group situational test where a candidate's ratings, at least to some extent, depend upon the particular combination of candidates with whom he happens to be grouped for assessment.

The effect on validity of criterion variations from discussion to discussion also suggests that increased effort must be directed toward training the observers to develop a standard frame of reference which transcends any given group discussion.

2. The LGD may be expected to demonstrate greater discriminability among those higher on criteria of leadership potential than among those lower on such external criteria.

3. LGD validity may be raised by increasing the standard deviations of discussion ratings. Aside from further training of the raters and more emphasis on forcing the raters to make greater discriminations, a number of ways may be suggested to increase the validity of the LGD.

a. The length of discussion time may be lengthened from 30 minutes to an hour or more with the expectation that greater stratification in status may occur—although no evidence is available to support this contention.

b. All the candidates can be coached briefly on how to be successful discussion leaders. Klubeck and Bass (9) have shown that while brief coaching raises significantly the LGD ratings of participants who are fairly successful initially without such training, such training does not alter the LGD ratings of those who have been found initially unable to emerge as discussion leaders. This would suggest that briefly coaching all participants would lead to a greater dispersion in the LGD behavior, although, of course, a long period of training might be expected to do otherwise.

4. Size appeared negatively related to validity. However, the relationship was too low to be anything but suggestive. Although groups varied only from 6 to 8 in size in these analyses, these variations accounted for 3 per cent of the variance. A study may be warranted of the relation between group size and validity similar to Bass and Norton's (4) analyses of the relation between size and reliability.

Similar Doolittle analyses were made to see the extent to which each of the other 7 variables contributed to the variance of the reliability of discussion ratings from group to group and the extent each of the other 7 variables contributed to the variations in efficiency from group to group. The two obtained mul-

multiple correlations were .59 and .56 respectively; however little further knowledge was added to understanding of the relationships among the variables than had been found by inspection of the correlation matrix.

### Summary

In order to determine the extent to which it was possible to account for variations in the reliability and validity of the leaderless group discussion, the means and standard deviations were computed for 8 variables along which LGD's vary. Also, a mean intercorrelation matrix was computed among these 8 variables.

The most important findings of this and related analyses were that:

1. The validity of the individual LGD varied greatly from discussion to discussion while the reliability of LGD ratings appeared quite stable.

2. Absolute ratings of leadership performance by LGD observers appeared accurately sensitive to variations from discussion to discussion in the outside leadership status of the participants.

3. Discussion observers' ratings agreed most closely when discussions were average in effectiveness rather than extremely effective or ineffective.

4. The validity of a given LGD was higher, the higher the outside leadership status of the participants in the discussion, the more stratified this status, and the more diverse the LGD ratings the observers were able to assign.

These results suggested a number of ways in which it might be possible to raise the validity of the LGD for assessing leadership potential.

Received March 3, 1952.

### References

1. Arbous, A. G., and Maree, J. Contribution of two group discussion techniques to a validated test battery. *Occup. Psychol.*, 1951, 25, 1-17.
2. Bass, B. M., and Coates, C. H. Forecasting officer potential using the leaderless group discussion. *J. abn. soc. Psychol.*, 1952, 47, 321-325.
3. Bass, B. M., and Norton, F. T. M. Group size and leaderless discussion. *J. appl. Psychol.*, 1951, 6, 397-400.
4. Bass, B. M., and White, O. Situational tests. III. Observers' ratings of leaderless group discussion participants as indicators of external leadership status. *Educ. Psychol. Measmt.*, 1951, 11, 355-361.
5. Carter, L., Haythorn, W. Meirowitz, B., and Lanzetta, J. The relation of categorization and ratings in the observation of group behavior. *Hum. Relat.*, 1951, 4, 239-253.
6. Doll, P. A. Validity of LGD assessment of unacquainted women. Unpublished Master's thesis, Louisiana State University: Baton Rouge, 1952.
7. Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1950.
8. Harris, H. *The group approach to leadership-testing*. London: Routledge and Kegan Paul, 1950.
9. Klubeck, S., and Bass, B. M. Differential effect of training on persons of different leadership status. *Hum. Relat.* (In press).
10. Landry, H. A., Krugman, M., and Wrightstone, J. W. *Validation study of the group oral interview test*. New York: Board of Education, 1951.
11. Mandell, M. Validation of group oral performance test. *Personnel Psychol.*, 1950, 3, 179-185.
12. Taft, R. Some correlates of the ability to make accurate social judgments. Ph.D. Dissertation, U. of California: Berkeley, 1950.
13. Vernon, P. E. The validation of Civil Service Selection Board Procedures. *Occup. Psychol.*, 1950, 24, 75-95.
14. Wurster, C. R., and Bass, B. M. Situational tests: IV. Validity of leaderless group discussions among strangers. *Educ. Psychol. Measmt.* (In press).

## Validity of the Strong Vocational Interest Blank Nursing Key

Leslie Navran

*San Francisco State College*

In a study done in 1947 (3), the Strong Vocational Interest Blank was administered to two groups of girls who were entering nursing training at Stanford University and San Jose State College. The Stanford group ( $N = 26$ ) had a mean score of 43.8 points on the Strong nursing scale. The mean score of the San Jose State group ( $N = 44$ ) was 40.1.

In 1949, a follow-up revealed that 59 of the 70 girls had completed the first two years of the three-year program. The mean nursing scale score of this surviving group was 42.0. According to the manual for the Strong Blank (5), a score of 41 is the dividing line between the B and B-plus letter grades. Thus, the girls entering the last year of training (who were considered by school officials as being almost certain to graduate)<sup>1</sup> had an average nursing scale score equal to only the 16th percentile of the standardization group. Twenty-six of them had scores in the B, B-minus, and C range. These results have important implications with respect to the validity of the nursing key and the use of the key in vocational counseling.

Previous reports have indicated that in the past the nursing key has been more useful. In 1939, Hilgard (1) found that "those with ratings on the Strong below 'A' in nursing showed little likelihood of completing the nurses' training course." Moreover, all the girls in her sample who scored below A had dropped out of training by the end of the first year. Roper (4) reported an average nursing scale score of 57.1 (in the A range) for 33 high school senior girls who were interested in nursing.

It is true, of course, that Strong's test does not purport to measure success in getting through school. Rather, it purports to measure the interests of women who have continued in nursing for a considerable period of

time after completion of training. Nevertheless, the Strong test *has* been used to counsel students prior to their entrance into training, and the contrast between present-day and past results with the nursing key indicates that vocational counselors should now interpret B and C scores on the nursing key with caution because the predictive value of such scores may have lessened materially.

In view of the small size of the samples used by Navran (3), it may be rash to state flatly that a revision of the nursing scale is needed. However, the following considerations lend support to the possibility that such is the case:

It has been necessary in recent years to revise some of the scales on the men's form of the Strong (2, 6), and where marked changes in the scale have resulted, they have been attributed to developments in the occupation itself which made for changes in the composition of the people engaged in the occupation. There is evidence that this may also be true of nursing. For one thing, partly as a function of World War II and the current Korean conflict, there has been a serious shortage of nurses. The effect of this has been to recruit heavily for the profession, and this may be bringing girls into nursing who differ from the standardization group, but who nonetheless can and will become nurses. This is another way of saying that nursing may be drawing from a wider segment of the general population in terms of measured interests than was formerly the case.

Related to this is the discrepancy in age between nursing trainees and the standardization group. Inspection of the Strong manual (5) reveals the standardization group to have been 34 years old, on the average, when tested in 1942. This means that girls presently graduating from high school and entering nursing training are approximately 15 years younger than the standardization group with whom they are being compared. This

<sup>1</sup> The writer has been informed that 24 of the 26 girls at Stanford completed their training successfully.

age difference may also be a factor making for different likes and dislikes in the present-day nursing trainees.

Finally, and perhaps most importantly, nursing itself has become more complex and proliferated. There is an increasing differentiation being made between the practical nurse and the professional nurse. Also, specialization in psychiatric nursing is growing more common, perhaps as a function of the growth and development of psychiatry and clinical psychology. It should be noted, too, that "the only revised scales which have differed appreciably from the old scales are the physician and psychologist scales."<sup>2</sup> Since these professional people with whom nurses are in close association have changed so greatly, it may be reasonable to hypothesize that nurse trainees who can get along well with them may also be quite different from their older and successful fellow-nurses. This is speculation, of course, but in view of the results reported above, it makes the adducing of more data extremely pertinent.

<sup>2</sup> Personal communication from the consulting editors of this journal.

## Summary

Evidence is presented which casts doubt on the validity of the present nursing key of the Strong Vocational Interest Blank. Factors which may account for this finding are discussed.

Received April 28, 1952.

## References

1. Hilgard, Josephine R. Strong Vocational Interest scores and completion of training in a school of nursing. *Psychol. Bull.*, 1939, 36, 646.
2. Kriedt, P. H. Vocational interests of psychologists. *J. appl. Psychol.*, 1949, 33, 482-488.
3. Navran, L. The Super-Roper technique as a measure of interest in nursing. *J. appl. Psychol.*, 1950, 34, 417-422.
4. Roper, Sylvia A. *A test of interest in nursing*. Unpublished Master's thesis, Clark University, 1940.
5. Strong, E. K., Jr. *Manual for Vocational Interest Blank for Women*. Stanford, California: Stanford University Press, 1945.
6. Strong, E. K., Jr. Vocational interests of accountants. *J. appl. Psychol.*, 1949, 33, 474-481.

## Individual Differences in Ability to Fake Vocational Interests

Ralph Garry

*Boston University*

The purpose of this study was to investigate individual differences in ability to fake vocational interests, to determine if reliable individual differences in faking ability existed, and if such differences could be related to vocational selection. Three separate trials were made, requesting college students, who had previously been administered the Strong Vocational Interest Blank under standard directions, to obtain as high a score as possible on certain of the occupational interest scales. Derived faking scores for the several scales were correlated to determine generality of ability to fake, and also the reliability of such faking.

### Background

The present study was initially begun in conjunction with the Medical Specialists Research Project, a joint undertaking of Stanford University and the Surgeon General of the U. S. Army, having for its purpose the development of test instruments designed to facilitate assignment and classification of doctors into residency training programs.<sup>1</sup> It was hoped that if reliable individual differences existed, they could be used in distinguishing candidates specializing in psychiatry from those in surgery, the assumption being that psychiatrists would show greater insight into attitudes and interests.

Although several attempts have been made to devise measures of "social intelligence," these so-called tests of "social intelligence" and "social judgment" have been little better than crude measures of intelligence (2, 6, 8). On the other hand, the results of several studies have suggested that the ability to fake scores in predetermined ways on preference-type tests showed promise as a possible measure in the area of social intelligence or psychological insight (1, 4).

<sup>1</sup> The author is indebted to Lloyd G. Humphreys under whose supervision the present study was executed.

Strong (7, p. 685) reports that "testees can deliberately obtain high occupational interest scores when they try." Benton and Kornhauser (1), interested in the use of the Strong Interest Blank in selection of medical school students, asked a group of 34 undergraduate college students (mainly social science majors) to fake as high a score on the physician scale as possible. The results corroborate Strong's finding that faking is possible. Of greater interest was an indication that all of the group did not gain, giving support to the premise that the ability to fake occurs in differing degrees.

Of the few earlier studies, one of the most relevant to the present was Steinmetz's (5). A total of 46 junior college students, directed to fake high scores on teacher-administrator scales on the Strong Blank, made significant gains over original scores. Intercorrelation between original score, faked score, intelligence and gains made in faking showed that both original and faked scores correlated with intelligence significantly greater than .00, but the difference between them was not statistically significant. The correlation between gain made and intelligence was significantly negative. Steinmetz infers from this negative correlation that intelligence makes little contribution to the obtained faking, apparently overlooking the extent to which the negative  $r$  is an artifact of method of determining gain; individuals with low initial scores have much greater possibilities for gains.

The extent of the relationship between intelligence and the ability to fake scores is critically important to a conclusion that faking ability represents "social judgment." An attempt to obtain a partial correlation coefficient between these two variables holding the relationship of each with initial score constant produced coefficients in excess of 1.00, suggesting inaccuracies in the data as presented.

In a recent study Jessen (3) found that parents' responses on answering the Kuder

Preference Record and Bell Adjustment Inventory as they thought their children would respond correlated .75 with child responses. These findings give rise to the question of how far such faking ability extends; that is, will a more general population show as high a degree of faking or is such faking ability limited to particular situations?

The purpose of this study, therefore, is to determine the degree, extent and reliability of the ability to fake scores on the Strong Vocational Interest Blank for Men.

### Population

A separate group was used on each of three trials. Group 1 consisted of 178 male, college undergraduates enrolled in a general psychology course. Groups 2 and 3 included 75 and 91 students of both sexes enrolled in educational psychology courses. The latter two groups were more heterogeneous with respect to age and vocational experience. For purposes of the tables in this report only the data for Groups 2 and 3 are presented.

### Procedure

1. The Strong Vocational Interest Blank for Men was administered using standard procedure to a group of sufficient size to provide sub-groups (successful and unsuccessful at faking) of fair size.

2. Biographical data were obtained for each subject using a multiple choice questionnaire. The majority of the questions asked for responses regarding education, work experience, hobbies, career choice and father's occupation. The format of the questionnaire permitted the classification of responses to any single question into a dichotomy for use with a biserial correlation coefficient.

3. A measure of intellectual ability was obtained. The best measure available was the academic grade-point averages of the subjects.

4. The Strong Blank was readministered with directions to fake high scores on designated scales. On the first trial the group was instructed to answer as they thought: (1) a carpenter would; and (2) as a physician would. Although the results generally coincided with results of the second and third

trials, it was evident that the carpenter scale had been a poor choice, apparently being too easily faked. The original mean score was -45, mean faked score was 115 with insufficient spread of scores to permit a test of differential faking ability. The low reliability of .39 (first versus last half corrected by Spearman-Brown formula) confirmed the doubts regarding the carpenter scale.

5. In the repetitions of the experiment, four scales were used instead of two, obtaining faking on one-half of each of the four scales in order to remain within reasonable time limits for testing. There are a sufficient number of weighted items on each half of the Interest Blank to provide an adequate measure of faking.

The four scales chosen for the second administration were physician, minister, lawyer and president of manufacturing concern. They were chosen because they had the lowest intercorrelations with physician scale, adequate reliability and, more important, because they represented the interest factors shown to be present in the Strong Vocational Interest Blank for Men in several factor analysis studies (7, p. 314 f).

6. The reliability of the faking on the second and third trials was determined by preparing scoring keys for use with the IBM test scoring machine which provided a reliability coefficient based on odd versus even responses. All items with plus weights were given plus one weights, and all with minus weights were given minus one weights.

7. In establishing a score for faking ability, it was apparent that a high score on the faked tests did not certify to high degree of faking ability, rather the ability to increase one's original score was the measure of faking ability. The problem was to obtain a relatively uncontaminated measure of gain. The high negative correlations obtained by Steinmetz (5) when gain made was compared to original score nullifies gain made as a measure of faking ability, for its magnitude is a function of initial standing. Two methods were tried with approximately equal results. The first, a ratio of gain made to gain possible was rejected because of the tendency of ratio scores to produce spurious correlations under

certain conditions. The faking score adopted was the difference between score obtained under faking directions and the score predicted on the basis of the correlation between original and faked scores. This difference represents a measure of faking ability, independent of original score, which may be correlated with similar differences obtained on the other scales.

Scatter diagrams were prepared to check the distribution of regressed faking scores<sup>2</sup> about the regression line for predicted scores for each of the four occupations. This was done as an empirical check of the assumption that the difference scores used in the preceding intercorrelations were randomly distributed about the regression line, and independent of the size of initial score. The distributions observed supported such an assumption.

8. In order to establish any generality of faking ability, it was necessary to account for any variance in the correlation between faking scores that is associated with intelligence, education or experience of vocational or avocational nature. Faking ability, if it exists as a psychological characteristic of any generality, should have some independence from the aforementioned factors. Biserial correlation coefficients were computed using the regressed scores on the physician-faking and president-faking because of their higher reliability and the ease with which the groups could be dichotomized as non-informed or informed about the occupation judging from such data obtained on biographical information blank.

<sup>2</sup> The term "faking score" is used to designate the score based upon the difference between the obtained faking score and the predicted (regressed) faking score.

Given such independence, the test for the presence of faking ability depended on low intercorrelations between initial scores, between initial and faked scores, but high correlations between faked scores. This would show that the rank order obtained on faking differed from that on initial scales, either between scales or within scales, thus indicating generality of faking ability.

9. A final step in the treatment of the data was an item analysis of responses made on the faking of the physician scale, using upper and lower 27% of Group 1 and Groups 2 and 3 combined.

### Results

The data obtained in this study confirm the reports of previous investigators regarding the extent of faking that is possible on pencil-and-paper tests of personality and interest. Groups of individuals, given instructions to fake high scores on the Strong Vocational Interest Blank for Men, are able to obtain significant increases in the group mean, although there are some individuals at all score levels who do not gain. The faking apparently is not correlated with intelligence, sex, or information about an occupation. The biserial correlation coefficients between faking score on physician scale and grade point average were  $-.18$  and  $.22$  (Groups 1 and 2); between faking score and sex were  $-.06$  and  $-.02$ , and for faking score and information were  $.02$  and  $.00$  (for president, manufacturing, and physician scales with data for Groups 2 and 3 combined).

Table 1 shows that consistent gains were made in the means on all scales, with the

Table 1  
Means and Standard Deviations of Original and Faked Raw Scores:  
Group 2 (N = 75) and Group 3 (N = 91)

Scale	Group 2				Group 3			
	Mean		SD		Mean		SD	
	Orig.	Faked	Orig.	Faked	Orig.	Faked	Orig.	Faked
President	8	22	18	19	7	24	18	19
Lawyer	2	25	20	10	1	20	10	12
Physician	- 1	61	35	36	- 3	66	40	31
Minister	-16	61	25	21	-17	65	28	21

standard deviations remaining fairly constant, except for the lawyer scale. The smallest gain in the means, that made on the president scale for Group 1, is significant at greater than the .001 level of confidence. On the whole, the similarity of the means and standard deviations indicates the comparability of the groups. This does not hold for the lawyer scale. The decrease of 10 raw score points on the standard deviation is significant above the .001 level of confidence. The most reasonable explanation for the decrease, and similarly that found with the carpenter scale on the first trial, is the low reliability of the faking scores.

It is possible that some scales are more easily faked by all members of a group, resulting in decreased variability under faking conditions. However, it should be noted that the observed decrease in variability is not associated with the scale's having a higher proportion of easily faked items, assuming the number of such items to be proportional to the number of large scoring weights. (It was observed in item analysis of faking that all members of the group choose the correct response for interests that are obviously related to a given occupation.) Under such circumstances the number of items upon which faking differences could obtain would be proportionately smaller, resulting in decreases in standard deviation. Both minister and physician scale have a greater proportion of large scoring weights than lawyer scale.

Reliability coefficients for each set of raw faking scores are presented in Table 2, along with estimates of the reliabilities of the regressed scores, based on the formula for reliability of a difference score. With the excep-

Table 2

Reliability of Raw Faking Score and Regressed Faking Score (estimated) for the Four Scales for Groups 2 and 3

Scale	Raw Faking Score	Regressed Score (estimated)
President	.87	.73
Lawyer	.55	.67
Physician	.89	.79
Minister	.78	.80

Table 3

Intercorrelation of Faking Scales for Groups 2 and 3

Scales Correlated	Group 2	Group 3	Mean $r$
President, mfg. concern, and lawyer	.28	.27	.28
President, mfg. concern, and physician	.34	.35	.34
President, mfg. concern, and minister	.22	-.11	.05
Lawyer and physician	-.05	.16	.06
Lawyer and minister	.10	.16	.13
Physician and minister	.26	-.02	.12

tion of the lawyer scale, the reliability of the faking scores is comparable to that reported by Strong (7) for scales administered under standard conditions.

Table 3 presents the correlations between the faking scores on the various scales. These indicate that there is no marked general faking ability; evidence, all  $r$ 's are under .36. The fact that nearly all  $r$ 's are positive, however, indicates the possibility of weak general faking ability, which could be proved or disproved in a subsequent trial by using scoring keys based on item analysis of responses of upper and lower faking groups. If true, one would expect increased correlations, using such keys.

The correlation between two faking scores is independent of the initial correlation between scales (as reported by Strong) judging from second order partial correlation coefficients computed between minister and president scales, which changed negligibly from the given .00  $r$  between original and faking scores.

Items analysis of faking responses to physician scale using top and bottom 27% of groups indicates that the differences obtained result from less than half of the weighted items. Both groups choose the obvious responses of physicians. Successful faking is dependent on predicting the more subtle differences in interests. Significant differences are obtained on as many unweighted as weighted items, suggesting considerable inaccuracy in faking. However, the differences obtained do not result from a willingness of

the successful faking group to commit themselves to a like or dislike response while the non-fakers remained neutral.

the subtle occupational interests, whereas all predict the obvious.

Received April 21, 1952.

### Summary

Two groups of 75 and 91 college students, instructed to fake high scores on four scales of the Strong Interest Blank after taking it under standard directions, demonstrated:

1. Significant increases in mean scores on all scales.

2. Split-half reliability of faking ranging from .56 to .89, with three scales exceeding .75, indicating a high degree of consistency.

3. Intercorrelations between faking scores (the difference between obtained and regressed fake score) ranging from  $-.05$  to  $.35$ , suggesting a low degree of general faking ability involved, with most faking being specific to the given scale.

4. Faking ability was not correlated (biserial  $r$ ) with intelligence, sex, or information regarding the occupation.

5. The more successful in faking (in an item analysis) predict substantially more of

### References

1. Benton, A. L., and Kornhauser, S. I. A study of "score-faking" on a medical interest test. *J. Ass. Amer. Med. Coll.*, 1948, 23, 57-60.
2. Broom, M. E. A further study of the validity of a test of social intelligence. *J. educ. Res.*, 1930, 33, 403-405.
3. Jessen, Margaret. *Parent-child cooperation in the counseling process*. Unpublished Ph.D. dissertation, School of Education, Stanford University, 1950.
4. Kelly, E. L., Miles, Catherine, and Terman, L. M. Ability to influence one's scores on a pencil-and-paper test of personality. *Char. and Personality*, 1936, 4, 206-215.
5. Steinmetz, J. C. Measuring ability to fake occupational interest. *J. appl. Psychol.*, 1932, 16, 123-130.
6. Strang, Ruth. Relation of social intelligence to certain other factors. *Sch. Soc.*, 1930, 32, 268-272.
7. Strong, E. K. *Vocational interests of men and women*. Stanford, Calif.: Stanford Univ. Press, 1943, xxix, pp. 746.
8. Thorndike, R. L. Factor analysis of social and abstract intelligence. *J. educ. Psychol.*, 1936, 27, 231-233.

## The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale

A. W. Bendig

*University of Pittsburgh*

One of the first problems faced by the constructor of a rating scale is the paucity of experimental literature on the optimal characteristics of such scales. Information is needed, for example, as to the effect of variations in the number of scale categories and in the amount of verbal definition or anchoring of the scale categories upon both the reliability and validity of the scales. The scale should not be so coarse as to lose some of the discriminative ability of the rater, nor so fine that error variance is added to the ratings because the scale categories call for finer judgments than the rater is capable of making. As to anchoring, presumably the more defining of scale categories and the more objective are such definitions the greater will be inter-rater measures of reliability. However, in self-ratings, such as are commonly used in personality studies (3), objective and extensive verbal definition of scale categories may result in an undesirable loss in the "projective" elements present in such self-ratings.

Two reports have discussed the effect of variations in number of scale categories upon reliability. Symonds (9) based a rational analysis of the problem upon Kelley's correction of an obtained correlation coefficient for coarseness of grouping in the measured variables (7, p. 168). He concludes that the reliability of the scale should increase as the number of scale categories increases, but that this increase in reliability is minor above nine categories. In view of the increased difficulty of the task for the rater, Symonds concludes that the optimal number of categories is from seven to nine. The empirical results of Champney and Marshall (4) question Symonds' analysis. In their study social workers rated visited families as to sociability on two forms of a graphic rating scale. These ratings were quantified by measuring the

graphic ratings using a millimeter scale and also using a coarser centimeter scale. The correlation between two forms of the scale (80 families rated twice) was significantly higher for the millimeter scale when compared with the centimeter scale (0.77 compared with 0.67). Such a magnitude of increase is much greater than could be predicted for Symonds' analysis. Bendig and Hughes (2) found that an information analysis of rating scales differing in number of categories indicates that the absolute amount of information transmitted by the scale increased with increasing numbers of categories, but that the increments became smaller with longer scales.

The purpose of the study reported below was to investigate the effect of variations in the number of scale categories and amount of verbal anchoring upon inter-judge (1) reliability estimates of self-ratings of individuals and of groups.

### Procedure

*Scales.* Fifteen different forms of a numerical rating scale were constructed from the combinations of five different numbers of scale categories (3, 5, 7, 9, or 11) and three conditions of verbal anchoring of the categories (center category defined, both end categories defined, or center and end categories defined). The lowest category on each scale was given a numerical value of 1, the highest category was rated as 3, 5, 7, 9, or 11, with intermediate scale categories numbered accordingly.

*Subjects.* The Ss were 225 undergraduate students in introductory and social psychology classes. The fifteen scales were randomly distributed among the subjects with 15 raters using each of the scales.

*Instructions.* Each scale was mimeographed on a single page containing the stim-

Table 1

Analysis of Variance of Group and Individual Reliability Coefficients of Rating Scales Differing in the Number of Scale Categories and Amount of Verbal Anchoring of the Scale

Source of Variation	df	Group Reliability			Individual Reliability		
		Sum of Squares	Mean Square	F	Sum of Squares	Mean Square	F
Total	44	4049.24			5105.64		
Number of categories	4	91.91	22.98	—	150.53	37.63	—
Amount of Anchoring	2	274.17	137.08	1.42	455.24	227.62	2.27
Interaction	8	793.91	99.24	1.03	1496.97	187.12	1.87
Within groups	30	2889.25	96.31		3002.90	100.10	

uli to be rated and instructions to the rater. The stimuli were the names of twelve foreign nations, ranging from well-known countries such as France and Canada to lesser known nations like Sweden and Egypt. The Ss were asked to rate themselves on how much they knew about the political, economic, geographic, and sociological characteristics of each country. Emphasis in the instructions was placed upon the Ss rating their own information about each country and the three verbal statements used to anchor scale categories were:

- I know a great deal about this country.
- I know something about this country.
- I know very little about this country.

### Results

Each group of 15 raters using one of the 15 different scales was randomly subdivided into three groups containing 5 raters each. An estimate of the reliability of group ratings for each subgroup was computed using the technique developed by Hoyt (6) and a similar estimate of the reliability of individual ratings was found using the intraclass method described by Snedecor (8, pp. 243-246) and elaborated upon by Ebel (5). The Hoyt procedure is designed to answer the question of the reliability of the *mean* ratings of five raters on the above described judgmental task and the Snedecor method estimates the reliability of a single rater on the same task. The resulting 45 group reliability coefficients were analyzed within the framework of a factorial design for the effect of variations in number of scale categories, amount of scale

anchoring, and the interaction of these two variables. A similar analysis of variance was computed on the 45 individual reliability coefficients. The results of these two analyses can be found in Table 1. It can be seen that neither of the two main variables contributed significantly to the total variability of either the group or the individual reliability coefficients. Also, in neither case was the interaction term significant when tested against the within-groups (error) mean square.

The three subgroups using each of the fifteen scales were pooled and new Hoyt and Snedecor reliability estimates computed. In this analysis each of the estimates is based upon the ratings of 15 subjects. Since the interaction of the main variables was insignificant, the three anchor groups were further pooled and group and individual reliability coefficients computed for each of number-of-scale-categories groups. These estimates are

Table 2

Average Group and Individual Reliability Coefficients (Decimal Points Omitted) for Each Number of Categories on Rating Scales

Type of Reliability	Number of Raters	Number of Scale Categories				
		3	5	7	9	11
Group	5	68	68	67	69	65
	15	89	88	87	89	84
	45	96	95	96	96	90
Individual	5	28	31	33	33	29
	15	34	34	32	35	27
	45	33	32	33	36	16

Table 3

Average Group and Individual Reliability Coefficients  
(Decimal Points Omitted) for Each Amount  
of Verbal Anchoring of Rating Scales

Type of Reliability	Number of Raters	Amount of Anchoring		
		Center	End	Center and End
Group	5	67	65	71
	15	86	87	89
Individual	5	29	28	35
	15	29	31	36

based upon 45 raters in each group. The average group and individual reliabilities for each of the category groups can be found in Table 2. In general, both the group and individual reliabilities were constant when 3, 5, 7, or 9 scale categories were used. However, in all instances the reliability declined somewhat when 11 scale categories were used and this decrease in reliability becomes more evident as the number of raters increases.

Similar average group and individual reliabilities for the three anchor groups are given in Table 3. Increased reliability can be noted with increased amounts of anchoring with the greatest increase occurring between the group with both ends anchored and the group with center and end anchoring.

Discussion

The results reported above suggest that the reliability of group or individual self-ratings is little affected by variations in the number of scale categories within the limits of from 3 to 9, but both individual and group reliability begins to decline when 11 categories are used. This conclusion is in opposition to what would be predicted by Symonds' analysis (9), but suggests that the point made by Champney and Marshall (4), i.e., that rating tasks beyond the discriminative ability of the rater adds error variance to the ratings, is confirmed in this instance. Presumably self-ratings using an eleven-category scale presents the rater with an introspective problem that is slightly too difficult and the reliability of his responses begins to decrease. While Champney and

Marshall found increased reliability with increased refinement of the scale, we have found opposite results. An explanation of this difference probably lies in the type of rating task presented to the subject. Champney and Marshall had their subjects rate the observed behavior of others: we had our subjects rate their own introspections. Obviously we cannot generalize our results to ratings of objective behavior, but must limit ourselves to the behavior herein investigated.

As to anchoring of the scale, increased verbal definition of the categories resulted in slightly increased reliability. The important anchor seemed to be that defining the center category. There was only a slight difference between the groups that had only the center category defined when compared with the group having only the two end categories anchored, but the addition of a center anchor to the latter scale appreciably raised its reliability. The lack of interaction between number of categories and amount of anchoring may be attributable to the fact that the categories added to the three-category scale were *unanchored* categories inserted between the center and end categories. Possibly longer scales, each of whose categories was verbally anchored, might not have shown a drop in reliability between 9 and 11 scale points.

Synthesizing these results with those previously reported (2) we recommend that in constructing self-rating scales 9 categories should be used, since: (a) they are as reliable as shorter scales; and (b) they provide more information. However, adding additional categories provides some increase in information at the sacrifice of scale reliability. It is further concluded that more verbal anchoring of the scale will increase both the reliability and the information transmitted by the scale.

Summary

A total of 225 college students rated themselves as to how much they knew about twelve foreign countries. The rating scales differed in number of scale categories (3, 5, 7, 9, or 11) and in amount of verbal anchoring of the scale points (center category defined, end categories defined, or both center and end defined). The reliabilities of individual and of group ratings

for each scale were computed by intraclass methods. Results indicated equal reliability for scales having 3, 5, 7, or 9 categories, but a decrease in reliability for 11 categories. The reliability of the scales increased with added scale anchoring. The discussion emphasizes that the results can be generalized only to self-ratings and not to ratings of observed behavior.

Received May 9, 1952.

#### References

1. Bendig, A. W. Inter-judge vs. intra-judge reliability in the order-of-merit method. *Amer. J. Psychol.*, 1952, 65, 84-88.
2. Bendig, A. W., and Hughes, J. B., II. The effect of the amount of verbal anchoring and number of rating scale categories upon transmitted information. (In preparation).
3. Cattell, R. B. The description of personality: principles and findings in a factor analysis. *Amer. J. Psychol.*, 1945, 58, 69-90.
4. Champney, H., and Marshall, Helen. Rater's minimal discrimination as a criterion for determining the optimal refinement of a rating scale. *J. appl. Psychol.*, 1939, 23, 323-331.
5. Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, 16, 407-424.
6. Hoyt, C. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
7. Kelley, T. L. *Statistical method*. New York: Macmillan, 1923.
8. Snedecor, G. W. *Statistical methods*. (4th ed.) Ames, Iowa: Iowa State College Press, 1946.
9. Symonds, P. M. On the loss of reliability in ratings due to coarseness of the scale. *J. exp. Psychol.*, 1924, 7, 456-461.

## An Analysis of Engineering Entrance Examinations

Harry W. Case

*Department of Engineering, University of California, Los Angeles*

The problem of measuring engineering achievement and aptitude is certainly not new. Indeed, this is an area of investigation that has been under study for the last twenty years. Today if the studies related to this field are compiled, they add up to an impressive number. In appraising engineering aptitude the exploratory investigations have ranged from determining the interrelationship existing between the scores of general capacity tests and measures of success in an engineering curriculum to attempts to tease out the specific factors making for success in engineering. For example, interrelationships as high as .62 (5) have been obtained between success in the first and second semesters and the American Council Psychological Examination (a general capacity measure).

When tests have been designed specifically to measure factors believed necessary to assure success in mastery of the engineering curriculum, the reported correlation coefficients are among those that are usually considered high for aptitude testing. In one study conducted in twelve schools the median correlation with first term grade averages and the Pre-Engineering Inventory was .60 (3).

At the College of Engineering, University of California, Los Angeles campus, a study of some of the variables influencing student success or failure has been underway since 1945. This engineering college is in many respects somewhat unique in the engineering educational field since it does not follow the usual sequence of undergraduate specialization in civil, electrical, and mechanical engineering but rather emphasizes the basic principles essential to all of these fields. One of the unusual features is the gradual incorporation into the program of the premise that engineering should utilize not only the laws of the physical sciences but also those derived from the life sciences (1).

### Subjects

Although the investigation of the factors that make for success in this engineering college has been underway for a number of years, many of the students who have entered and proceeded either to graduation or withdrawal could not be used as subjects in this study. The elimination of numerous cases,—such as those in which courses were repeated to raise a grade, or in which previous specialized and related military training existed, pre-entrance engineering extension division study had been taken, and other related and influencing extraneous variables,—greatly reduced the total number of cases available for study. From a total of well over a thousand potential subjects the actual correlations were obtained for  $N$ 's which ranged from 144 to 444. Even though the reduction in the total number of subjects available for the measurement of the various interrelationships is regrettable, it is believed that matching the subjects in terms of previous training more than offsets the loss of mass data.

It is probably somewhat unfortunate that in the majority of studies published in the last ten years no mention has been made as to whether the subjects have been matched in terms of previous training and preparation, although it is recognized that many students who enter as freshmen have had prior college or military training which may influence their success in the first two academic years.

### Procedure

All incoming freshmen were given the complete Pre-Engineering Inventory prior to entrance. This is a special abilities test battery using the "task-simulation" technique and was developed as a joint project of the Engineers' Council for Professional Development, The American Society of Engineering Education, and the Carnegie Foundation for the Advancement of Teaching. It consists of the seven

Table 1

Tetrachoric Intercorrelations\* of P.E.I., Jr. Status Examinations, Certain Subject Areas, and Semester Grades

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
PEI Gen. Verb.	PEI Tech. Verb.	PEI Sci. Mat.	PEI Gen. Math.	PEI Mech. Prin.	PEI Spa. Vis.	PEI Mod. Soc.	PEI Composite	PEI Total	Jr. Status—Engl.	Jr. Status—Draw.	Jr. Status—Chem.	Jr. Status—Math.	Jr. Status—Phys.	Jr. Status—Total	Chem. Grades	Math. Grades	Physics Grades	Drawing Grades	Lib. Arts H. S.	Pre-Engr. H. S.	1st Sem. Aver.	2nd Sem. Aver.	3rd Sem. Aver.	4th Sem. Aver.
1	62	67	30	28	17	64	64	72	75	06	26	23	34	43	36	19	30	23	02	03	01	23	31	42
2		67	60	45	25	50	86	80	35	30	48	22	42	56	41	26	31	02	02	08	24	44	39	56
3			65	65	55	58	91	88	54	25	45	31	53	72	38	39	46	22	03	11	23	39	43	49
4				62	41	40	85	71	35	19	46	47	44	48	40	29	26	16	44	31	20	41	46	46
5					51	41	66	73	32	17	39	21	53	49	32	24	40	11	05	02	13	28	25	31
6						21	47	55	30	46	37	25	25	50	26	09	07	28	20	39	18	20	23	23
7							55	70	45	31	34	06	26	44	39	13	38	-10	-02	12	17	33	31	39
8								92	49	30	57	21	60	69	50	40	44	24	12	04	32	54	54	54
9									67	42	64	23	61	77	50	38	44	23	09	11	46	51	48	50
10										29	37	33	35	61	46	41	28	23	27	20	04	28	30	33
11											38	11	34	41	29	22	20	41	28	33	28	31	32	35
12												32	49	77	65	40	49	15	27	31	39	49	53	48
13													51	61	38	47	34	13	37	35	14	29	37	37
14														73	52	30	44	09	39	32	32	49	51	54
15															56	46	52	24	24	23	31	55	74	65
16																71	80	43	47	38	15	63	70	66
17																	86	49	26	38	54	72	83	85
18																		53	19	28	59	74	85	90
19																			25	22	20	22	45	45
20																				78	20	32	32	33
21																					42	43	42	39
22																						61	55	45
23																							62	63
24																								65

\*  $r$ 's in boldface are significant at the 1% level.

following named tests: General Verbal Ability, Technical Verbal Ability, Comprehension of Scientific Materials, General Mathematical Ability, Comprehension of Mechanical Principles, Spatial Visualizing Ability, and Understanding of Modern Society. An examination of the material of each section of the test indicates that its face validity closely approximates the name.

If a student was moving from sophomore to junior status, either within the college or by means of a transfer from a junior college, he was required to take the Junior Status Examination. This is an achievement examination battery of the multiple choice type consisting of five separate tests, each covering a

specific field: Chemistry, Physics, Mathematics, English, and Drawing. The mathematics and drawing tests are University of California, College of Engineering Examinations, while the other three are from the Cooperative Test Series, Higher and College Level.

In addition to these examinations, the student's previous high school grade record was used for evaluation and admission. For admission purposes the high school record was divided into two categories: those subjects which were classed as liberal arts and those subjects which were termed pre-engineering. The pre-engineering group of courses consisted of mathematics, physical sciences (chemistry and physics), mechanical drawing, and Eng-

lish. The courses remaining after these were deducted from the transcript were loosely classed as liberal arts.

Tetrachoric  $r$ 's and the Standard Error<sup>1</sup> for each of these  $r$ 's were calculated between the entrance devices described above and the first four semesters of work as well as the grouped subject areas of chemistry, mathematics, physics, and drawing. These correlations, as well as the internal interrelationships, are shown in Table 1.

### Criteria

The question of the reliability of specific subject grades as well as the reliability of the semesters' grades needs to be considered, because if the reliability is low, little can be done to increase the validity of a test for selection purposes. The high intercorrelations existing between chemistry and physics grades .80, chemistry and mathematics grades .71, and physics and mathematics grades .86, would seem to indicate that some reliability exists, since the material learned in these three courses is related. Similarly, the fairly consistent intercorrelations between the first four semesters would appear to substantiate this belief. If the intercorrelations between semesters may be taken as an index of reliability, the reliability may be said to be as high as the relationship obtained between the measuring instruments and the criteria, i.e., tests and high school grades versus college semester and subject grades.

### Results

An examination of the correlations in Table 1 reveals some interesting trends. One of the first immediately noticeable is the magnitude of the intercorrelations existing between the sections of the P.E.I. (Pre-Engineering Inventory), when these are compared with the  $r$ 's existing for the various sections of the Junior Status Examination. One possible explanation for this difference is that it is easier to

<sup>1</sup> The SE's were estimated by applying the formula:

$$SE_{r_i} = \frac{1.5 (1 - r^2)}{\sqrt{N - 1}}$$

According to Garrett (2), "An approximation to the SE of a tetrachoric  $r$  may be found in the following way: the  $r$  is about 50% higher than the SE of an equivalent product-moment  $r$ . . . ."

segregate the subject areas covered in an achievement type examination into measurable units with little overlap. The substantial overlap between sections of the P.E.I. arouses a question as to the success with which it will predict the grades for a total semester and as to the differential value of its various sections for specific subject areas.

Both the P.E.I. Total score and Composite score predict the four semesters' grades fairly well, the one exception being a .32 correlation between the first semester's work and the P.E.I. Composite score. The other correlations between the semesters' grades and the P.E.I. Total and Composite scores range from .46 to .54, which is close to the median of .60 reported by Johnson for twelve engineering colleges (3). On the other hand, the various sections of the P.E.I. show fairly low correlations with grades for specific subject areas. The highest single correlation existing between a subject area and a section of the P.E.I. was .46 for the P.E.I. Scientific Materials and Physics grades. This same lack of relationship and inability to discriminate between subject areas was found by Moredock (4). It would appear, therefore, that while the test shows usefulness in predicting over-all success for the first two years of engineering curriculum it is of little use in evaluating potential student success for specific subjects.

At this point it is perhaps desirable to note that the P.E.I. Total and Composite scores show little relationship with grades obtained in high school "pre-engineering" subjects, or with grades obtained in those high school subjects which have been designated as "liberal arts." This low intercorrelation, combined with the fact that the "pre-engineering" high school subjects relate at a median of .42 with the first four semesters of college work, has allowed the P.E.I. Total score and the "pre-engineering" high school grade scores to be combined into a successful selection device. It would appear desirable eventually to design a new examination which would show less overlap between its sections and greater differential value for subject areas.

In an analysis of the results of the Junior Status Examination, which is of the achievement type, it will be seen that the intercor-

relations for its sections range from .11 to .51. These low intercorrelations may in turn be responsible for the higher  $r$ 's of the examination with the specific subject areas. The differential value is highest for chemistry, mathematics, and drawing. The physics section of the examination, which has its highest correlation with chemistry grades, was a shortened version restricted to Mechanics and Electricity. The intercorrelations between both "pre-engineering" and "liberal arts" high school subject grades with the total score of the Junior Status Examination are also greater than those obtained with the P.E.I. The correlations of the examination and the first four semesters of work range from .31 to .74.

Two additional correlations which are not included in Table I have been obtained for the total of the Junior Status Examination and success in the fifth and sixth semesters of work. The correlation between the fifth semester of work and the examination is .65, and the sixth semester is .58. These two  $r$ 's were obtained by the Pearson product moment method and are based upon 100 and 68 cases respectively.

It should perhaps be noted that while this paper has been devoted to an analysis of the interrelationships existing between the results of the examinations and high school grades when used for entrance evaluation and the grades received in the first two years of engineering college, the examinations have proved useful in many other ways that are difficult to quantify. For example, information obtained from one of the examinations has been used in conjunction with a diagnostic interview to determine the areas in which remedial work is needed.

### Conclusions

1. The Pre-Engineering Inventory shows a consistent correlation with the grades from the first four semesters of work, which makes it useful as a selection device.
2. The sections of the Pre-Engineering Inventory show no clearly defined relationship with specific subject areas, which would make it useful for differential selection within engineering.
3. The Pre-Engineering Inventory shows a low interrelationship with "pre-engineering" subject high school grades.
4. High school "pre-engineering" subject grades show a consistent correlation with grades from the first four semesters of engineering college work.
5. An achievement examination such as the Junior Status Examination shows both greater differential value and greater over-all relationship with semester grades.

Received March 10, 1952.

### References

1. Case, H. W. The utilization of psychology in engineering courses. *Amer. Psychologist*, 1951, 9, 494.
2. Garrett, H. E. *Statistics in psychology and education*. New York: Longmans, Green and Co., 1947.
3. Johnson, A. P. Tests and testing programs. *J. Engng. Educ.*, 1951, 41, 277-282.
4. Moredock, H. S. *Special abilities in pre-engineering studies*. Unpublished Doctor's dissertation, Univ. Calif., 1950.
5. Remmers, H. H., and Geiger, H. E. Predicting success and failure of engineering students in the school of engineering in Purdue University. *Purdue Univ. Stud. Higher Educ.*, 1940, 38, 10-19.

## Differential Sex Responses to Items of the MMPI

L. E. Drake

*Student Counseling Center, University of Wisconsin*

In a rather extensive study of the MMPI some incidental evidence has come forth which appears important enough to warrant an early report. The frequency of Yes, No, and ? responses to each of the 550 items of the card form was obtained separately for 2,270 undergraduate male students and for 1,148 unmarried, undergraduate female students.<sup>1</sup> All were enrolled in the University at the time of testing and none obtained an L score over 70 or an F score over 80.

Excluding items to which 90 per cent or more of both groups responded in the same direction and excluding those to which 10 per cent or less responded in the same direction, 306 items yielded critical ratios between the sexes ranging from 2.0 to 32.3. A total of 43 items was selected from these 306. These were items that 50 per cent or more of the females

responded to in a direction in which less than 50 per cent of the males responded. Answer sheets for 100 males and 99 females *not* included in the original groups were scored for the 306 items and the 43 items. The resulting coefficient of correlation was  $+.80$ .

Answer sheets for 3,229 males and 1,612 females were then scored with the 43-item key. Only 2% of the females obtained a score as small as or smaller than the mean score for the males and only 2% of the males obtained a score as large as or larger than the mean score for the females. That this sex difference is reliable is further indicated by the fact that a coefficient of correlation of  $+.80$  was obtained between scores (43-item scale) obtained by 474 males and 224 females on the group form taken at time of entrance to the University and the card form taken up to one year later.

It is quite apparent that sex is an important factor in establishing criterion groups, especially for scale construction for this type of inventory.

*Received October 23, 1952.*

*Early publication.*

<sup>1</sup> Tables of frequency counts for each item by sex have been deposited with the American Documentation Institute. Order Document 3860 from American Documentation Institute, c/o Library of Congress, Washington 25, D. C., remitting \$1.25 for photocopies (6 × 8 inches) readable without optical aid or \$1.25 for microfilm (images one inch high on standard 35 mm. motion picture film).

## A Study of Medical Students with the MMPI: III. Personality and Academic Success

William Schofield

University of Minnesota

Two previous papers in this series have reported general normative data for samples of medical students studied with the Minnesota Multiphasic Personality Inventory (1) and the nature of changes in the MMPI profiles of students from the freshman to junior years of the medical curriculum of the University of Minnesota (2). This paper, the last in the series, is concerned with the relationships between MMPI profiles and academic success.

### Class Standing and MMPI Profile

The Dean of the Medical Sciences provided data on the total honor point ratios at the completion of the junior year of the members of the class used in this study. The Student Counseling Bureau provided the American Council on Education Psychological Examination (ACE) scores of the students. With these data at hand, it was possible to select students from the upper and lower quarters of the class who were matched for scholastic aptitude as measured by the ACE. These matched groups were then studied for similarities and differences on the MMPI.

The forcing of homogeneity on the aptitude variable resulted in very small samples (11 students each) from the upper and lower quarters, but this was considered preferable to the use of larger N's with uncontrolled aptitude variance. The differences between the ACE scores of matched upper and lower quarter students ranged from zero to nine percentile points, with an average difference of five percentile points.

Figure 1 shows the mean freshman year profiles of the upper and lower quarter samples. Three of the clinical scales reveal a statistically reliable difference in the mean scores of the two groups. The lower quarter group had reliably higher mean scores on the Hy, Pd, and Sc scales. Table 1 presents the data for these comparisons. The differences

between the two groups are seen to be limited to a very few scales and, while statistically reliable, are not great. In general, as seen in Figure 1, the profiles of the upper and lower quarter students are similar, particularly in terms of the relative elevations of the "char-

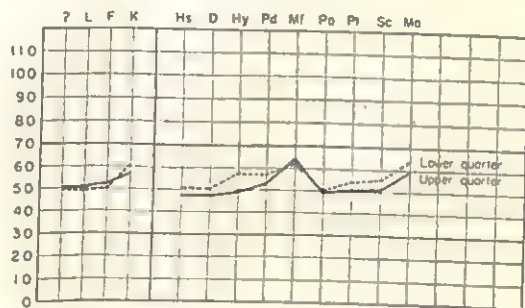


FIG. 1. Mean freshman year MMPI profiles of samples of medical students from the top and bottom quarters of their class at the end of the junior year. (N = 11).

acter structure" scales of the right side of the profile. To the degree that the good and poor achievers are distinguished by the MMPI, the distinction appears in the relative degree of hysteroid, psychopathic, and schizoid tendencies; the poorer achievers are characterized by a tendency to unrealistic appraisals of their environment, unhappy social relationships, and autistic rumination. Also, the poor achievers show a general tendency toward a relatively unsophisticated denial of personal weaknesses and the expression of an idealized self-concept (L). In this regard, the students who work up to capacity tend to manifest a more realistic self appraisal.

As an approach to testing the predictive significance of the group differentiations turned up in this comparison of mean profiles, the total class of students was sorted into two groups: (1) a group each member of which had an MMPI profile characterized by one or both of the two highest scores falling on the Hy, Pd, or Sc scales; and (2) a group whose

Table 1

Means and Standard Deviations of the Freshman MMPI Scores for Samples of Upper Quarter and Lower Quarter Medical Students Matched for Scholastic Aptitude<sup>1</sup> (N = 11)

Scale	Group				t	F
	Upper Quarter		Lower Quarter			
	Mean	Sigma	Mean	Sigma		
L <sup>2</sup>	1.2	1.4	2.7	1.4	4.26**	1.10
F <sup>2</sup>	3.7	3.2	3.2	2.6	.46	1.50
K	57.1	5.9	61.0	7.4	1.48	1.55
Hs	47.4	7.8	50.9	7.1	1.03	1.19
D	47.4	9.5	50.5	8.8	.65	1.16
Hy	49.3	7.0	57.4	6.8	3.08*	1.06
Pd	53.0	6.7	57.1	10.1	3.03*	2.29
Mf	64.0	6.9	62.0	13.4	.36	3.74
Pa	49.7	8.6	50.0	7.2	.10	1.44
Pt	50.9	6.3	54.6	9.0	1.09	2.07
Sc	51.0	5.2	55.4	7.4	2.06*	2.01
Ma	59.7	10.0	63.5	7.3	1.28	1.88

<sup>1</sup> Academic standing determined from honor point ratio at end of junior year. Upper and lower quarter students matched for ACE.

<sup>2</sup> Statistics based on raw score data; for raw scores on L which are less than 3, arbitrarily T scores of 50 are set.

\* Significant at 5% level.

\*\* Significant at 1% level.

profiles did not have the above characteristics. From these two groups, two smaller samples were drawn so that each member from the group with Hy, Pd, or Sc high points was matched with a member from the other group for ACE score. Then a study was made of the honor point ratios of these two samples which were equated for scholastic aptitude but differentiated by the presence and absence of certain scales as high points in their MMPI profiles. Table 2 reports the mean honor point ratios (HPR) of these two samples. The group characterized by profiles with high points on the Hy, Pd, or Sc scales yielded a mean HPR clearly inferior to that of the group not so characterized. However, since the variances of the two groups differed significantly, it was not possible to test for the reliability of the difference between the means. It may be concluded, nevertheless, that these samples do not support the hypothesis that the populations of which they are representative have identical distributions of honor point ratios.

Figure 2 shows the actual distribution of HPR's for the two samples. The greater

range of achievement in the group not characterized by high points on Hy, Pd, or Sc is clear from this figure. While there is clear overlap of the two distributions, a cutting line at HPR = 1.6 shows only 23% of the group with the specified high points to have HPR's larger than this value, while 62% of the other

Table 2

Means and Standard Deviations of the Honor Point Ratios of Two Groups of Medical Students Differentiated by MMPI High Points and Matched for ACE Scores

Group <sup>a</sup>	N	Honor Point Ratios <sup>b</sup>		ACE %ile	
		Mean	Sigma	Mean	Sigma
A	21	1.48	.22	64.7	22.8
B	21	1.80	.36	63.9	21.9

F = 2.68\*

<sup>a</sup> Group A had profiles for which one or both of the two highest scores fell on the Hy, Pd, or Sc scales. Group B did not have either of their two highest scores on the Hy, Pd, or Sc scales.

<sup>b</sup> Total honor point ratio at end of junior year.  
\* Significant at 5% level.

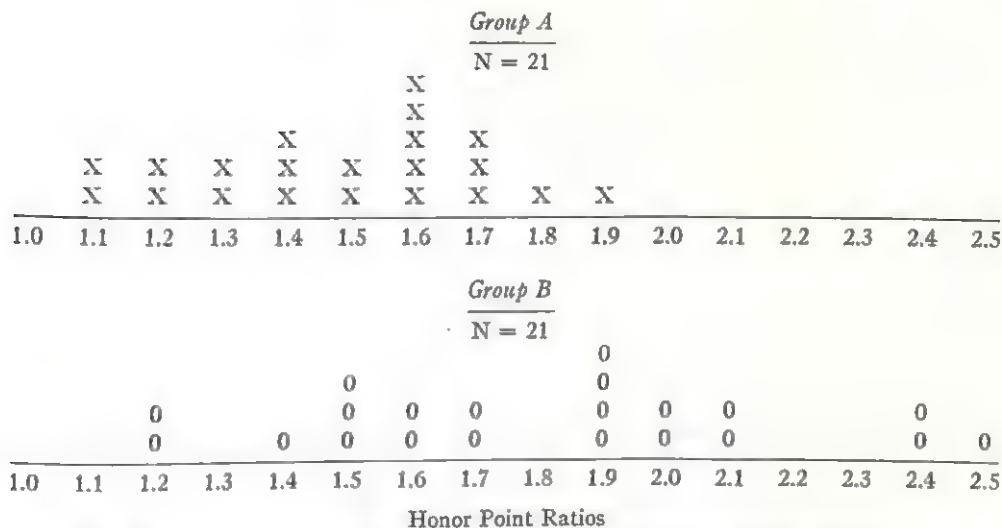


FIG. 2. Distribution of honor point ratios of two groups of medical students differentiated by MMPI high points and matched for ACE scores.\*

\* Group A had profiles for which one or both of the two highest scores fell on the Hy, Pd, or Sc scales. Group B did not have either of their two highest scores on the Hy, Pd, or Sc scales.

group fall above  $HPR = 1.6$ . At the other end of the distribution, while 43% of the Hy-Pd-Sc high point group have  $HPR's < 1.5$ , only 14% of the other group fall below this score. It appears that the presence of the highest or next to highest score of a medical student's profile on the Hy, Pd, or Sc scales is highly predictive of underachievement.

Another approach to this study of relationships between personality and medical school achievement was made by studying the relative success of students with deviant MMPI profiles and those with profiles within the normal range. There were 18 members of the class (21.6%) who had freshman MMPI profiles with at least one of the clinical scales showing a T-score of 70 or greater. Fifteen of these constituted the "deviant" sample. Fifteen students with profiles entirely within the normal range were selected so as to be matched with the "deviant" group for ACE scores. The range of the differences between the ACE scores of the matched students ran from zero to six percentile points. Table 3 reports the mean honor point ratios for the "deviant" and "non-deviant" groups at the end of the junior year. Figure 3 shows the mean profiles of the "deviant" and "non-deviant" group. The mean profile of the "deviant" group reflects the fact that ten of

the fifteen students in this group had a score of 70 or greater on the *Mf* scale. It is obvious that the two groups are essentially identical in their academic performance as expressed by the honor point ratio.

It was considered of interest to make one additional study of MMPI profiles and achievement. This was a study of the relationship between medical school class rank at the end of the junior year and the amount of difference between the freshman and junior year MMPI profiles. For this purpose the same matched samples of upper and lower quarter students for whom mean honor point ratios are reported above were used. Figures 4 and 5 indicate the freshman and junior profiles of these two samples. It is quite clear that the upper quarter students show a much

Table 3

Means and Standard Deviations of Honor Point Ratios (Junior) of Medical Students with "Deviant" and "Non-Deviant" Freshman MMPI Profiles

Group	N	ACE Percentile		Honor Point Ratio	
		Mean	S.D.	Mean	S.D.
Deviant	15	73.4	5.48	1.58	.49
Non-Deviant	15	72.7	5.23	1.56	.32

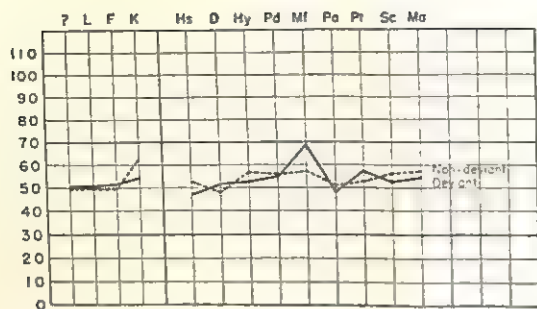


FIG. 3. Mean MMPI profile of a sample of fifteen medical students having at least one clinical score over  $T = 70$ , and mean profile of fifteen students with no deviant score, both samples matched for ACE scores.

greater change in their mean profile over the two year interval than do the lower quarter students. The top quarter sample showed a reliable increase in mean score on the Sc scale, and statistically significant decreases in means on the Mf and Ma scales. Thus, the top quarter students showed a tendency after two years in the medical curriculum toward a "defeminization" of their interest and activity pattern (Mf) although remaining clearly deviant from the general population males. Likewise, their morale, optimism, enthusiasm and self confidence showed a drop toward the general population norm (Ma) which may reflect a more realistic appreciation of their capacities and the demands of medical training. The increase in Sc suggests a tendency to greater self analysis and general philosophical probing which is probably in line with the drop in manic features.

By contrast, the bottom quarter sample revealed little tendency to reliable change over

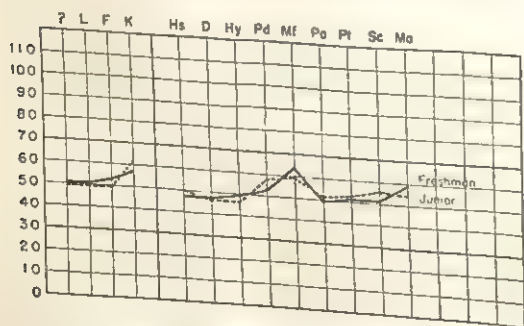


FIG. 4. Mean freshman and junior year MMPI profiles of a sample of medical students in the top quarter of their class at the end of the junior year. ( $N = 11$ ).

the two year interval, the sole change having statistical reliability being a drop in the Ma score suggestive of mild deterioration of morale. Tables 4 and 5 present the means and standard deviations of the scale scores for both years and both samples together with measures of the reliability of the freshman-junior differences.

As a further check on the relationship between amount of change in MMPI profile and academic performance a scattergram was prepared to show the joint distribution of these two variables for the entire class of 83 students. The "change" score for each subject was obtained by adding, without regard to sign, the differences between his freshman and junior scores on each of the nine clinical scales. For the total sample, this variable showed a range of 29–118 T-score points, with a mean of 54.90 and a standard deviation of 17.16.

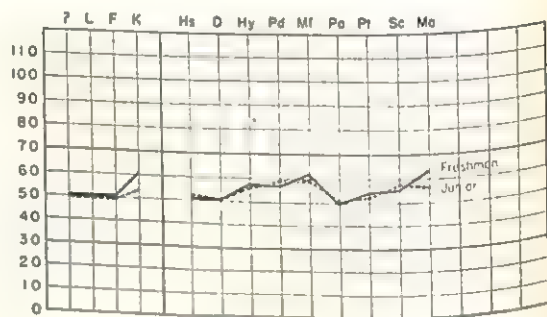


FIG. 5. Mean freshman and junior year MMPI profiles of a sample of medical students in the bottom quarter of their class at the end of the junior year. ( $N = 11$ ).

The honor point ratios for the group had a range of 1.1–2.8, with a mean of 1.71 and a sigma of .35. Inspection of the scattergram did not suggest any marked relationship between these two variables although it did appear that there was a slight tendency for higher honor point ratios to be associated with higher change scores. Table 6 indicates the means and sigmas of the honor point ratios for the subjects having the 20 lowest and the 20 highest MMPI change scores. The difference between the mean honor point ratios of these two groups is not statistically reliable. Comparison of the distribution of honor point ratios for these two groups revealed considerable overlap.

Table 4

Means and Standard Deviations of Freshman and Junior Year MMPI Scores for a Sample of Top Quarter Medical Students (N = 11)

Scale	Freshman		Junior		t	F
	Mean	Standard Deviation	Mean	Standard Deviation		
L	1.2	1.4	1.5	1.1	.42	1.59
F	3.7	2.8	3.1	1.8	.98	2.50
K	57.1	5.4	61.6	7.2	1.66	1.79
Hs	47.4	7.1	50.0	5.6	1.06	1.57
D	47.4	8.6	46.8	8.0	.26	1.16
Hy	49.3	6.4	46.2	7.7	1.38	1.45
Pd	53.0	6.1	57.6	9.6	1.16	2.47
Mf	64.0	6.2	59.3	7.4	2.65*	1.41
Pa	49.7	7.8	51.3	6.0	.68	1.68
Pt	50.9	5.7	52.3	6.4	.56	1.25
Sc	51.0	4.7	56.0	5.7	3.52**	1.47
Ma	59.7	9.2	54.8	7.8	2.22*	1.38

\* Significant at 5% level.

\*\* Significant at 1% level.

Table 5

Means and Standard Deviations of Freshman and Junior Year MMPI Scores for a Sample of Bottom Quarter Medical Students (N = 11)

Scale	Freshman		Junior		t	F
	Mean	Standard Deviation	Mean	Standard Deviation		
L	2.7	1.2	1.8	2.4		3.50*
F	3.2	2.4	2.4	1.4		2.98*
K	61.0	6.8	53.0	6.2	.85	1.18
Hs	50.9	6.4	51.7	5.1	.03	1.60
D	50.5	8.0	50.2	7.4	.01	1.15
Hy	57.4	6.2	56.5	5.6	.07	1.24
Pd	57.1	9.2	58.6	7.0	.53	1.72
Mf	62.0	12.2	59.2	9.6	.94	1.61
Pa	50.0	6.5	50.2	5.9	.13	1.21
Pt	54.6	8.2	52.7	5.1	.57	2.59
Sc	55.4	6.6	57.5	6.8	.70	1.05
Ma	63.5	6.6	56.6	7.1	2.50**	1.14

\* Significant at 5% level.

\*\* Significant at 1% level.

### Summary and Conclusions

Using total honor point ratio at the end of the junior year of medical school as a criterion, an attempt was made to investigate the relationship between personality tendencies, as revealed in a freshman year MMPI profile, and academic performance. Also a study was

made of the relationship between amount of personality change between the freshman and junior years and scholastic achievement. These analyses were based on data for 83 male students who entered the University of Minnesota Medical School in 1946.

1. When the average profile of upper quar-

Table 6

Means and Standard Deviations of Honor Point Ratios of the 20 Students Having the Lowest Amount of MMPI Score Change, Freshman to Junior Year, and of the 20 Students with Greatest Change

Group	N	Honor Point Ratio		
		Mean	Sigma	
Low Change	20	1.7	.28	F = 1.746
High Change	20	1.8	.37	t = .84

ter students was compared with that of lower quarter students, with the subjects of the two samples matched for ACE scores, certain of the scales revealed reliable mean differences between the two groups. The scales yielding reliable differences between the top and bottom quarter samples were Hy, Pd, and Sc. In general, the low quarter students revealed a tendency toward greater neuroticism and defection in interpersonal and social relationships.

2. When students were separated into two groups with members of the groups matched for academic aptitude (ACE) but differentiated by the occurrence and non-occurrence of high points of their profiles on the Hy, Pd, or Sc scales of the MMPI, it was found that the group having such high points was clearly inferior in academic performance (HPR) to the group not showing these scales as high points. Ninety per cent of the group with high points on Hy, Pd, or Sc had honor point ratios falling below the median HPR (1.75) of the group without these high points.

3. The fact of an MMPI profile with at least one elevated clinical score ( $T \geq 70$ ) did not appear to be predictive of inferior academic performance. When a group of students each of whom had at least one elevated score was compared with a group having no elevations, with members of the two groups matched for ACE, the honor point ratios of the two groups were found to be essentially identical.

4. It was found that the samples of first and fourth quarter students, equated for ACE scores, were very different with respect to the amount of change in their respective MMPI profiles from the freshman to the junior years. The top quarter students revealed a reliable change in mean score on three of the nine clinical scales (Mf, Sc, and Ma). The bottom quarter sample showed reliable freshman-to-junior changes only in decrease in their Ma score.

5. When a comparison was made of the honor point ratios of students having the largest amount of change in their clinical MMPI scores from the freshman to the junior years and the honor point ratios of students with the smallest amount of change, it was found that the two groups had essentially identical honor point ratios and there was considerable overlap between the two groups.

6. In general, it appears that when academic aptitude is constant, the likelihood of achievement up to capacity in the medical curriculum becomes less as hysteroid, psychopathic, and schizoid traits, measured by the MMPI, are greater. It may be hypothesized that students who show *both* a restricted scholastic promise and marked deviation on the Hy, Pd, or Sc scales would be particularly poor academic risks. In the absence of any limitation of academic aptitude, the admission to medical training of students showing chief deviations (even though within the "normal" limits) on the Hy, Pd, and Sc variables would appear to make for a lowering of the general level of scholarship of the medical school class.

Received April 28, 1952.

### References

1. Schofield, W. A study of medical students with the MMPI: I. Scale norms and profile patterns. *J. Psychol.*, 1953, 36, in press.
2. Schofield, W. A study of medical students with the MMPI: II. Group and individual changes after two years. (In press.)

## A Note on Ranking Method

Douglas Irvine

*Army Operational Research Group, Surrey, England*

Kendall (2, p. 89) mentions: ". . . the desirability of examining the primary data to see if there are any obvious effects present." The present note aims to enlarge upon this. Ranking is often a quick and meaningful method of obtaining various types of psychological data. Sometimes subjects are asked to rank various items in order of preference, such as those in the Job Preferences Scale of Jurgensen (1). Usually, in order to obtain some sort of over-all picture such rankings for each item are summated, and mean ranks are then calculated. After this the mean, or total scores are placed in rank order once more.

Such a procedure may conceal important information. This is becoming apparent in data being accumulated on an Anglicized version of Jurgensen's Scale, but, as the number of subjects in this study is small, the data will not be published at present. It seems that such data should first be arranged to show the number of times each item is placed in each rank. In other words a frequency distribution should be made for each item. This distribution may be turned into a graph for those who prefer to look at their results in this way.

The following example should make this clear, although it is much more simple than is usually encountered, since preferences are asked concerning three items only. Wyatt, Langdon, and Stock (3, p. 12) asked 19 operatives engaged on chocolate (candy) packing to rank in order of preference three sizes of boxes, viz., 14 lb., 4 lb., and 1 lb. (The situation is somewhat unreal in Britain today.) Their results are as follows:

	1st	2nd	3rd
Large boxes	10	1	8
Medium boxes	3	14	2
Small boxes	6	4	9

If these figures are summated, their mean ranks calculated, and the boxes arranged to give a final ranking for all operatives, we have the following results:

	$\Sigma x$	$\bar{x}$	Rank
Large boxes	36	1.89	1
Medium boxes	37	1.95	2
Small boxes	41	2.16	3

This latter table suggests that there is little difference in the over-all order of preference, whereas in fact the operatives tend to either like, or dislike, both the large and the small boxes, according to individual choice, while most of them place the medium boxes in the middle. The investigators found this to be of some importance, as there was a close correspondence between output in packing one type of box and preference for that box. This, however, is not the place for an argument into which is cause and which effect. The example is merely an illustration of Kendall's plea.

*Received April 24, 1952.*

### References

1. Jurgensen, C. E. Selected factors which influence job preferences. *J. appl. Psychol.*, 1947, 31, 553-563.
2. Kendall, M. G. *Rank correlation methods*. London: Chas. Griffin and Co., Ltd., 1948, pp. 160.
3. Wyatt, S., Langdon, J. N., and Stock, F. G. L. *Fatigue and boredom in repetitive work*. I.H.R.B. Report No. 77, London, H.M.S.O., pp. 86.

## Identification of American, British, and Lebanese Cigarettes

E. Terry Prothro

*American University of Beirut, Lebanese Republic*

Habitual smokers generally believe that they can differentiate between various brands of cigarettes. In many countries a smoker has a wide choice of both domestic and foreign cigarettes, and there is often a substantial difference in price between one brand and another. Obviously smokers must believe in a discriminable superiority of the more expensive brands if these brands are to be smoked for reasons other than "conspicuous consumption." Advertisers, of course, encourage the belief that different cigarettes have unique characteristics.

Investigations, however, by American psychologists throw considerable doubt on the belief that cigarettes can be identified by persons who do not know the brand they are smoking. Hull (1) found in the course of investigations on another problem that his Ss frequently failed to distinguish between tobacco smoke and warm moist air if visual cues were eliminated by a blindfold. Husband and Godfrey (2) requested blindfolded Ss to identify five American brands of cigarettes, and found that performance was only slightly better than chance on all brands except a mentholated one.

More recently Ramond, Rachal and Marks (3) examined the ability of habitual smokers to identify three popular American brands. They gave each of their subjects a practice smoking session during which there was opportunity to study the characteristics of the three brands. They did not blindfold their subjects on grounds that "a blindfold obscures the central problem." Thus the subjects could examine the texture of cigarette paper, the color and size of the tobacco shreds, etc. During the test session gummed labels were placed over the brand names. Their subjects were able to identify each of the three brands slightly more often than chance. Smokers who preferred one of the three brands were able to identify that brand significantly

more frequently than could smokers of other brands.

If we grant that even habitual smokers in America have difficulty in distinguishing between American brands of cigarettes, two questions present themselves. Is the difficulty a result of similarity of all tobacco smoke? Does the fact that American subjects in these experiments tend to smoke one brand of cigarettes to the exclusion of others affect their ability to identify non-preferred brands?

The situation in Lebanon is well suited to a preliminary investigation of these questions. Both American and non-American brands are used extensively, and the difference in price of various brands causes college students to vary the brand purchased as their own financial status fluctuates. Also there is some fluctuation in availability of brands on the market.

In Lebanon, as in most of the Arab Near East, American, English and domestic cigarettes are available. Of these the American cigarettes are the most expensive. English brands cost about 10 per cent less. Domestic cigarettes—made from tobacco grown in the Near East—are about half as expensive as American brands. The sale of cigarettes is under control of a government-supported monopoly which establishes prices and determines what cigarettes are to be imported. At the present time two American brands, Camel and Lucky Strike, and two English brands, Players and Gold Flake, are found on the market.

### Procedure

Subjects were 50 male college students who stated that they smoked at least five cigarettes per day. They were obtained by asking for volunteers from the student body of the American University of Beirut.

Each S was brought into a well-ventilated room and shown a table on which there were six packages of cigarettes. There was one

package each of the following brands: Camel, Lucky Strike, Gold Flake, Players, Bafra and Star. The choice of American and English cigarettes was based on availability to consumers. Bafra and Star were selected as being among the most popular of the Lebanese cigarettes. An effort was made by *E* to use cigarettes of equal freshness. The *S* was told that he would be presented with one cigarette from each of the six packages in turn, and that he was to try to identify each cigarette immediately after smoking it. He was warned that each guess was final and that he could not change his opinion about one cigarette after smoking some of the others. Thus if he guessed Bafra for the first cigarette and then decided that the second cigarette was actually the Bafra he was permitted to name the second "Bafra" but not permitted to change his guess on the first cigarette.

The *S* was next asked which cigarette he preferred. Then he was seated and blindfolded. Cigarettes were placed in wooden holders which were 6 cm. long. The holder was placed in the *S*'s mouth and the cigarette was lit for him. He was not permitted to touch or to see the cigarette at any time. As soon as he identified the cigarette it was removed and placed in a water-filled can. The *S* was then permitted to rinse his mouth at a water fountain just outside the experimental room. Approximately two minutes elapsed from the time one cigarette was identified until the next one was lit. It was hoped that the use of blindfolds and holders would minimize

available cues so that successful identifications not attributable to chance might be attributed to the qualities of the smoke itself.

The order in which the cigarettes were presented varied from subject to subject, and was determined by use of a table of random numbers.

### Results

It can be seen from Table 1 that our subjects were able to identify the American and English brands about half of the time and to identify the Lebanese brands even more often. Bafra was the most easily identified of these brands. Only seven of the subjects failed to identify it. Of the 300 attempts at identification, 180 or 60 per cent were correct. These results are considerably better than chance, and the superiority to chance is highly significant statistically. The value of chi-square for Table 1 is 462. This value is much too large to be found in the average table of chi-square. Moreover, each brand was identified better than chance. If we consider only the cells which pertain to correct identification, we find that the value of chi-square for these cells varies from 32 to 122. All of these values are highly significant.

From these results it appears that all cigarette smoke is not the same. Habitual smokers can differentiate between these six brands.

The difference between our results and the conclusions of Husband and Godfrey, and of Ramond *et al.* might lead us to conclude that the cigarettes which are popular in the Near

Table 1  
Number of Subjects Giving Each Response after Smoking Each of the Brands

Brand Smoked	Brand Named by Subject					
	Camel	Lucky Strike	Gold Flake	Players	Bafra	Star
Camel	24	5	3	12	3	3
Lucky Strike	8	28	8	4	1	1
Gold Flake	7	7	25	8	1	2
Players	5	6	8	27	0	4
Bafra	0	1	1	2	43	3
Star	4	3	2	0	8	33
Total	48	50	47	53	56	46

East are less similar to each other than are the popular American brands. There is, however, another possible explanation. Lebanese students vary the brands smoked to a greater extent than do American students. Consequently the Lebanese students may be better able to differentiate cigarettes because of a more varied experience.

In this connection it should be noted that our Ss identified the American brands quite successfully. There was little confusion between Camels and Lucky Strikes.

The results of Ramond *et al.* support the thesis that preference determines identifiability in America. Their Ss could identify the brand which they preferred more than 70 per cent of the time. On the other hand, those Ss who preferred a brand other than the ones used in the experiment averaged only 20 per cent correct identification, although chance performance was 33 per cent.

Of our Ss, 44 expressed a preference for one of the six brands and 26 (nearly six-tenths) of these were able to identify the preferred brand. When we recall that exactly six-tenths of all 300 trials were correct, it is apparent that our students could identify non-preferred brands as readily as they could identify preferred brands.

#### Summary

A total of 50 male students at the American University of Beirut who smoke at least five cigarettes per day were asked to discriminate between six brands of cigarettes which are popular in the Near East. Of the six brands,

two were American, two British and two Lebanese. Ss were blindfolded and presented with six cigarettes in succession. They were required to guess at the identification of each brand before proceeding to the next. All cigarettes were presented in wooden holders.

Ss were able to identify each of the six brands significantly more often than chance. Of all attempts at identification, 60 per cent were correct. It therefore appears that habitual smokers can discriminate between these cigarettes on a basis of the smoke alone.

It was pointed out that the superior performance of our subjects, even at distinguishing between Camels and Lucky Strikes, might be attributed to the tendency of Lebanese smokers to vary the brand smoked to a greater extent than do American smokers. The results are compatible with this thesis, for our subjects were able to identify non-preferred brands as readily as preferred brands. In contrast, a recent study (3) of American smokers demonstrated that they could identify the brand they preferred, but could not identify other brands.

*Received March 5, 1952.*

#### References

1. Hull, C. L. The influence of tobacco smoking on mental and motor efficiency. *Psychol. Monographs*, 1924, 33, 161.
2. Husband, R. W., and Godfrey, J. An experimental study of cigarette identification. *J. appl. Psychol.*, 1934, 18, 220-223.
3. Ramond, C. K., Rachal, L. H., and Marks, M. R. Brand discrimination among cigarette smokers. *J. appl. Psychol.*, 1950, 34, 282-284.



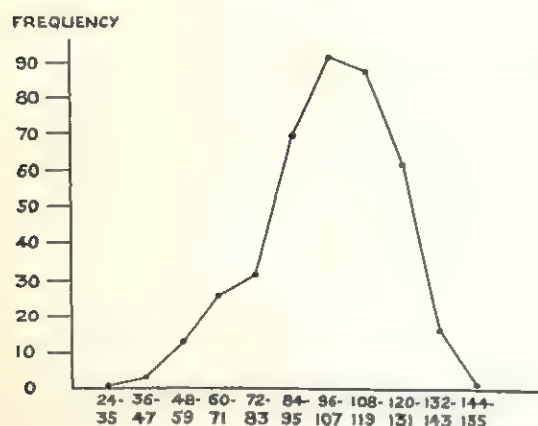


FIG. 2. Distribution of 400 standard scores on the stasiometer.

ing that hand operational steadiness and hand static steadiness are not related.

Another group of 50 subjects was given the Purdue Pegboard dexterity test. The correlation of these scores with those on the stasiometer was .057.

The stasiometer test was given to 50 skilled workmen (tool and dye makers, machinists,

Table 2

The Effects of Smoking and Sex on Operational Hand Steadiness

	N	Mean Steadiness Z Score	S.D.	C.R.
Smokers	225	101.9	20.3	
Non Smokers	175	99.2	21.5	1.27
Male	222	105.1	19.2	
Female	178	95.1	21.3	4.9

sheet metal workers, and welding inspectors). The average Z score for this group was 121.5, S.D. 16.7, as compared with a mean of 100 and S.D. of 30 for the norm group. The CR of the difference between these averages was 7.6. Further validity data are being collected.

Apparently the stasiometer is a reliable instrument for measuring operational steadiness and it may have some usefulness in selecting apprentices for various skilled occupations.

Received April 17, 1952.

## Book Reviews

Miller, Delbert C., and Form, William H. *Industrial sociology; An introduction to the sociology of work relations*. New York: Harper & Brothers, 1951. Pp. 896. \$6.00.

The objective of this book is to present the sociology of work relations. The word industrial is used as referring to all forms of economic activity. *Industrial Sociology* includes the study of occupations and all social groups that affect work behavior. Conceiving the subject from this point of view, the book deals with the interrelationships between the work behavior of the individual and the other aspects of his social activities. "The Framework of Industrial Sociology" (p. 30), as the five major subdivisions of the book, include: "I. Industrial Sociology: Its Rise and Scope; II. The Social Organization of the Work Plant; III. Major Problems of Applied Industrial Sociology; IV. The Social Adjustment of the Worker; and V. Industry, Community and Society."

Involving themselves with such broad interest areas without reaching encyclopedically precise detail will stimulate instructors and critics to judge the content and choice of the authors' evaluative selections. Such selections are exemplified by the identifying of "The rise of industrial sociology . . ." (p. 3), with the familiar Hawthorne experiments; a sample curriculum for the training of an industrial sociologist (p. 86), and a chart, included both on the front cover and on page 11, listing the chronological "Outlines of the Main Streams & Tributaries of Industrial Relations Knowledge Contributed by the Basic & Applied Social Sciences." The authors have provided an interesting base from which to work irrespective of the specific selections of the materials presented. Of particular value to this area which overlaps many disciplines is a glossary at the end of the book.

In a volume of this size there is much material which any particular group of readers may consider extraneous. The first 306 pages generally deal with basic informative material. To students who have had some particularized education in labor movements, industrial eco-

nomics, business administration, and courses in applied psychology, these materials may prove to be repetitious and tend to cause a general letdown before such students get to the more strictly sociological material. As was suggested in an article by the reviewer in *The Journal of Educational Sociology*, November 1950, ". . . the basic principles underlying industrial sociology are composed of established sociological principles and that industrial sociology represents a distinctive area of investigation for the sociologist which in large part he has left to economists and psychologists."

This volume is very comprehensive. As a reference for allied courses the comprehensiveness of this volume has great value in pointing up the interrelationship of the associated aspects of industrial relations. As a text for industrial sociology this very comprehensiveness makes it somewhat difficult to point up the basic underlying principles of this interest area. However, the book is unquestionably a real contribution to the role and teaching of industrial sociology.

Glaister A. Elmer

*Air University Far East Research Group*

*IES Lighting Handbook*. Second Edition. New York: Illuminating Engineering Society, 1952. Pp. 974. \$8.00.

This new edition represents a thorough revision of the original Handbook which was published in 1947. Its objective is to provide its readers with essential information on light and lighting in simple terms and condensed style. The introductory chapters are concerned with the physics of light, light and vision, and nomenclature together with definitions and symbols. These are followed with several chapters dealing with measurement of light, color, light control, daylighting, light sources and lighting calculations. There are then several chapters dealing with lighting in various situations such as interiors, exteriors, highways, aviation, transportation, and photography. The book is concluded with an extensive appendix. manufacturer's data (ad-

vertising) and an index. The numerous charts and illustrations are very useful and excellently done.

Although designed for use by illuminating engineers, there is much material included that can be useful to the applied psychologist. Special mention might be made of the sections on light and vision, nomenclature and measurement. Much of these materials should be known by the psychologist who is dealing with illumination in relation to visual comfort and efficiency. Other sections of particular interest to psychologists are the chapters on color and on interior lighting. Materials on recommended and standard practices are not included but the bulletins in these areas are listed opposite the title page.

The collection and organization of materials in this Handbook represents an extensive and difficult task. The committee in charge is to be congratulated on achieving an excellent result. No illuminating engineer or psychologist interested in the applied aspects of lighting can afford to be without this reference book. Nevertheless, there are a few reservations that occur to the reviewer: (1) There is a tendency to neglect psychological factors in adjustment of the individual to the illumination of working and living environments. In future revisions it might be well to include a chapter on this subject. (2) Considerable work in the field of illumination has been done by psychologists. Examination of the lists of references fails to disclose these reports except for rare instances. It would seem that the best results in lighting could be achieved by coordinating the work of engineers with that of psychologists, physiologists including medical men, and physicists. (3) The presentation of certain data may lead to misinterpretations. For instance, in presenting Weston's data, curves for relative performance but not for actual performance are given. The uncritical reader might interpret the curves presented to mean that, if the illumination is high enough discrimination of the low contrast test object will equal that of the high contrast one. Examination of the performance curves in the original report reveals that this is not so. In a similar manner, the data on speed of vision (Cobb, Ferree and Rand) is plotted in terms

of the reciprocal of the time. This produces an exaggerated picture of the improvement with increase in illumination intensity.

Miles A. Tinker

*University of Minnesota*

Frederiksen, N., and Schrader, W. B. *Adjustment to college*. Princeton: Educational Testing Service, 1951. Pp. xvii + 504.

Based on a study of 10,000 men veteran and non-veteran students in sixteen American colleges following World War II, this book has much to contribute to current educational and psychological theory and practice. Even though the population studied may, it is hoped, never be duplicated, the extent and form of this investigation are such that its implications are and will be important.

The book has a somewhat novel organization, for the whole study is summarized in the first chapter on a level clearly appropriate for the statistically untrained reader. In the remaining chapters, the results are presented in generally simple tabular and graphic form, while the basic tables and methodological notes are contained in the appendices. Although the first chapter is clearly intended for the lay reader, the level of difficulty of the rest fluctuates somewhat more than would seem desirable, and the college administrator who is tempted to read on may encounter rough going in certain places.

Two methodological points are of special interest. The authors use as the criterion of academic adjustment an index called the "Average Adjusted Grade," a "... measure of achievement-relative-to-ability. . . ." It is a standard score based on analysis of covariance procedures and represents a significant advance over other similar methods of computing such indices. Secondly, they make use of a sign test in assessing group differences, recognizing that samples from several colleges may be taken as replications of the experimental situation. This test makes for a precision too seldom encountered; it is hoped that it and the above criterion method will receive the attention they deserve.

The content of the book is a detailed description of the attitudes and behaviors of a group of men veterans and non-veterans and

a comparison between these. In the process, the authors dispose of a number of erroneous notions about these groups in particular and college students in general. For one thing, the similarities reported are more evident than are the differences, and the authors wisely recognize the importance of such "negative" results. Consequently, the book contains many descriptions of generally applicable relationships—and lacks of relationship—between the criterion and such factors as extra-curricular activities, vocational decision, family income, outside reading habits, etc.

Most of the book deals with the results of an extensive questionnaire intended to illuminate causes of obtained differences, if any. The authors are aware of the limitations of this method, and they report only what the students *said*. But it is easy to accept such statements at face value, something which might be, in view of what is known about test-taking attitudes, seriously misleading. It is to be hoped that other investigators will follow the many interesting leads provided and conduct studies employing more powerful tools.

For any complete picture of the problem of college adjustment, there is a great need for the integration of such studies with those by men such as Pressey whose work on educational acceleration makes a fairly clear case in favor of the younger collegian. However, since this book is not intended to be a systematic integration but rather an extensive descriptive study, this should not be taken as a criticism of it but only as an indication of a pressing need for more work.

Methodologically, this study should serve as a model for further research. As far as the results are concerned, both college administrators and psychologists interested in human adjustment problems should find in it a very great deal that is of interest and value. The study was imaginatively planned and carefully executed; the authors are to be commended for their excellent contribution.

John W. Gustad

*University of Maryland*

Kelly, E. L., and Fiske, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: The University of Michigan Press, 1951. Pp. 311. \$5.00.

This volume is the report of an ambitious five-year research program during the period 1946–1951, which was directed at the evaluation of techniques for the selection of graduate students for training in a four-year doctoral program in clinical psychology.

The first section of the report is devoted to a description of the operating philosophy of the project, which was to be both catholic and eclectic in the selection of predictors and criteria, a discussion of the sequential phases of the research program, and a presentation of normative data descriptive of the 700 subjects. Each subject was enrolled in one of 40 universities and had field training in one of 50 VA installations. As one would anticipate, the normative data indicated that there was a hierarchy of universities in terms of ability and achievement of their students, and there were large differences in emphases of training programs at the various universities and VA installations.

The second section deals with the three types of predictor measures under study. The first of these was a group of predictions by university staff members of the success of entering students upon examination of credentials only and upon examination of credentials plus interview in the following areas: Academic Performance, Skill in Diagnosis and Therapy, Research Competence, and Overall Promise as a Clinical Psychologist. The second type of predictor was a series of objective tests from which 101 measures were obtained. These objective tests were commonly used measures of intelligence, interest, and personality, among the specific tests being the Miller Analogies (Form G), the Strong Vocational Interest Blank, and the Minnesota Multiphasic Personality Inventory. The third type of predictor measure was a series of ratings based upon clinical procedures, which included intensive interviews and projective tests, and both individual and pooled ratings. Most interesting was a description of a pilot assessment program in which group situational tests

were utilized in addition to other techniques. Factor analysis was performed to identify the first-order factors of the some 42 variables under investigation in the pilot assessment program.

The third and largest section describes the development of criterion measures. There are interesting analyses of many problems encountered, such as the first-order versus second-order and specific versus general criteria problems. The authors found no satisfactory single criterion of success, although they did identify three general components of success. These were: intellectual accomplishment, clinical skills of diagnosis and therapy, and skills in social relations. They found judges agreed much better on the first than the other two components. There are many ideas and findings from which others concerned with similar searches for the criterion will-o'-the-wisp may profit. Of course, the criteria developed are in a sense also predictors of later performance as clinical psychologists. Until a follow-up study has been made and the criteria utilized in this program have been related to on-the-job success, the findings of the program are questionable. The authors do state that they hope to follow-up their subjects some ten, fifteen or twenty years later. Certainly this study does merit a fitting sequel.

The fourth section presents data upon the degree to which the predictor measures correlate with the criterion measures and contains a thorough discussion of various factors which have an influence upon the magnitude of the correlation coefficients.

The final section contains a summary of the major findings. Space prohibits the reviewer from commenting upon most of the findings presented either in this section or throughout the volume, which is literally studded with interesting findings. To the reviewer it was most significant that single objective tests predicted most of the criterion measures (including global measures, such as "Rated Overall Clinical Competence") just about as well as more laborious and time consuming ratings by professional staff members, and that single projective tests were almost worthless in predicting criterion measures.

In addition, there are several appendices

which present many of the devices utilized in the study and certain other important information, such as rejected criterion measures.

While the general aim of the program as stated in the Preface was to evaluate techniques for the selection of professional personnel, the authors do not purport to resolve all problems even within the limited area of the selection of clinical psychologists. Certainly most of the predictive findings cannot be generalized to the selection of personnel for training in other professional areas, although many of the techniques should offer valuable suggestions to researchers. However, it is concentrated attacks of this nature which should eventually lead to the improvement of the selection of personnel for training in the professions. The study is *must* reading for all those working in the areas of prediction of professional success and of criterion research.

Stanley E. Jacobs

Department of the Army,  
Washington, D. C.

Parker, W. E., and Kleemeier, R. W. *Human relations in supervision*. New York: McGraw-Hill, 1951. Pp. vii + 472. \$4.50.

At one time or another, most personnel men have struggled with the problem of improving the human relations skills of company supervisory personnel so there is a great deal of interest in any text which may prove to be useful in discussion or conference groups concerned with handling human relations problems.

In the authors' words, *Human Relations in Supervision* is "directed specifically to the first-line supervisor, because the establishment of good human relations in any organization stands or falls upon the skill of these supervisors in dealing with human problems." In general, the authors have succeeded in keeping the material at this level, employing many anecdotes, illustrations, and case studies in their attempt to relate human relations principles to the everyday experience of supervisors. One undesirable outcome of this level of treatment, however, is that much of the discussion on topics such as motivation, counseling, leadership and personal development is superficial.

In directing this book to the first-line supervisor, the authors have emphasized the handling of problems originating with the employee and do not treat directly the problems of the supervisor and his impact on the work group, the relationships between the various supervisory and management levels, or the effects of company policy and organization structure on the supervisor.

At least two suggestions for improvement come to mind. First, an introductory section or, possibly, a separate manual outlining the experience in companies using this material together with a statement as to the instructional methods employed and the outcomes of the training would be of great value. Second, the discussion questions following each chapter should be reworked since in their present form they invite class members to parrot back text material.

In summary, the authors have done a good job in assembling materials for a human relations course for men at supervisory levels. Whether the textbook-classroom approach which is indicated can or will produce the desired change is still an unanswered question.

William E. Kendall

*The Chesapeake and Ohio Railway Company*

Gray, J. Stanley. *Psychology in industry*. New York: McGraw-Hill Book Company, Inc., 1952. Pp. vii + 401. \$5.00.

This book reflects the author's belief that any factor which affects the production efforts of workers is appropriately classified as industrial psychology. This view has resulted in a different type of book on psychology in industry. It is, however, a disappointing book.

Considerable emphasis is given human engineering, work curves, physical and physiological measurements of work, fatigue, efficiency, nutrition, rest, monotony, boredom, lighting and ventilation. Some subjects are handled differently than is customary; for example, merit rating is discussed in a chapter on wages. A five-page appendix describes and illustrates calculations of the mean, standard deviation, standard error of the mean, correlation coefficient, and significance of differences between means.

Although all subjects discussed in the book may legitimately be included in the field of industrial psychology, relative emphasis deviates sharply from that found in actual practice. For example, twenty pages, or five per cent of the entire book, are devoted to nutrition. Subjects which are usually emphasized are discussed only briefly; for example, employment interviewing is handled on one page. Thus the book should not be interpreted as giving a true picture of the field as it is commonly conceived.

The book has a number of faults: broad statements are undocumented, superficial definitions are used, "obviousness" is used to support statements, flat statements are made which run counter to experimental evidence published elsewhere, broad coverage of subject matter results in superficiality. On the other side of the ledger are favorable factors such as inclusion of material not generally readily available to beginning students, uncommon use of common sense, and astute insights. Unfortunately, however, the assets do not appear to offset the limitations of the book.

Clifford E. Jurgensen

*Minneapolis Gas Company*

Zaleznik, A. *Foreman training in a growing enterprise*. Boston: Harvard Business School, 1951. Pp. 232. \$3.50.

"Is [supervisory] training realistic from the supervisor's point of view and in relation to his problems at work? The only way to develop an answer to this question in a particular organization is to go to the work level, and to observe what is happening" (p. 232). The author had done just this. This book is concerned with the evaluation of a foreman training program in a small manufacturing firm through 5 weeks of intensive on-the-job study of one of the trainees, a newly appointed foreman. Two other approaches to evaluation are also reported—observation of the training sessions and interviews with foremen.

The training course evaluated appears to be a rather confusing hodge podge of academic psychology, rules-of-thumb for handling people, and pep talks—all of which are not uncommon approaches to foreman training in

American industry today. In terms of being of value to this particular foreman in equipping him to better perform his job, the course was unsuccessful.

Despite the rather shaky design upon which this study is built where conclusions are drawn and recommendations made based on an N of a single foreman, this book makes a contribution. Its main value is in the convincing and meaningful manner in which the many complex relationships with which a modern factory foreman must deal are described. Pointing up how inadequately a typical packaged training program fulfilled the on-the-job needs of the foreman only helps to accentuate and sharpen the picture of the complexity of his job.

There are a number of weaknesses in the study. The author tends to draw too many definite conclusions and to overgeneralize from his single case. Many of the conclusions and recommendations are colored by the background and training of the author. For example, the only recommended method of foreman training discussed in any detail is the case method. Recommendations on the kind of training which would have helped the foreman more, including such things as coaching by his superior and permissive rather than authoritative conferences, are not new. However, despite these shortcomings, against the background of the real needs of a live supervisor on the job, the conclusions and recommendations still are much more convincing than they are when they appear as mere statements of opinion as is usually the case.

Because no serious reader can come away from this book without a fuller realization of the problem faced in developing supervisors, it can be highly recommended to persons concerned with supervisory training. If the book is read by this group, the future ought to produce more of the broad and continuing

type of training needed for helping the foreman perform his difficult job.

Theodore R. Lindbom

*Midland Cooperative Wholesale,  
Minneapolis, Minnesota*

Welch, J. S., and Stone, C. H. *How to build a merchandise knowledge test*. Research and Technical Report 8, Industrial Relations Center, University of Minnesota. Dubuque, Ia.: Wm. C. Brown Company, 1951. Pp. 21. \$1.00.

This excellent monograph concisely presents the methods for the development of job knowledge tests. Although the purpose is to describe the steps in the construction of information tests for use in evaluating experience of salespersons, the procedures are general and can be applied to any type of job.

The authors do not claim that they are describing any new methods. What they have done is to bring together for trade tests the procedures for item development, item validation, test validation, cross validation, and the setting of critical scores, in a most clear and logical fashion. The rationale for each step is well outlined. The monograph is liberally documented with judiciously chosen illustrations, so that each step is readily understandable.

The monograph will not only serve as a technical manual for those concerned with selection problems, but should be an invaluable piece of outside reading for a course in test construction or in psychological measurement. The only shortcoming is in the discussion of the types of items that might be used in job information tests. While a reader unfamiliar with the field will ultimately obtain some notion concerning the scope of possible items, in no single section is this aspect well developed.

Edwin E. Ghiselli

*University of California, Berkeley*

# Journal of Applied Psychology

VOL. 37, No. 2

APRIL, 1953

## Function Analysis of Thirty-Two American Corporate Boards \*

Jerome G. Kunnath and Willard A. Kerr

*Illinois Institute of Technology*

The insight of the general public and even of industrial psychologists into the typical activities of corporate boards of directors is probably somewhat vague and inaccurate. Since the corporate board is an important policy-making nerve center in industrial society, it needs to be brought within the orbit of psychological research.

This study, profiting from the activity analyses reported by Flanagan on laboratory personnel (1), Gordon on airline pilots (2), and Wagner on dentists (3), is, however, focussed on group rather than individual behavior. What does the corporate board do at its meetings? What are some of the probable determinants of what it does? It is the purpose of this study to investigate some of the behaviors of the corporate board.

### Experimental Design

Invitations to participate in a nation-wide study of corporate board activities were sent to one board member of each of 246 corporations. These 246 names were selected at random from the "Corporation Directors and Executives, 1950." A total of 32 firms actually participated. In each instance a member of the firm's board of directors completed an "Industrial Board Member Survey" chart which listed 21 topics of board activity under the following heading: According to my business board experience, boards consider at how many meetings per year? The board member then indicated the number, out of twelve, of meetings per year at which each topic is considered.

\* Work completed in the student research program of the *Industrial Psychology Laboratory* of the Illinois Institute of Technology.

In size, the corporations sampled ranged from 50 to 25,000 personnel, the median being 250. Sizes of boards ranged from 2 to 23 members, 5 being the median. Mean ages of members of the 32 boards ranged from 49 to 73, the median being 58.7. The average number of other corporate boards to which the average board member in these firms belongs ranges from 0 to 14, the median being 2.6. The per cent of board members who also work in the operating management of a firm ranges from 13.3 to 100.0, the median being 66.6. Fourteen of the firms studied were in metropolitan areas (500,000 population or more), and 18 were in less populated localities. Geographic distribution of the 32 firms was closely representative of the national distribution of American industry. When classified according to kind of industry, the breakdown of firms is as follows: heavy, 7; heavy-light manufacturing and transportation, 12; light manufacturing, 8; commercial (retailing, utilities), 3; finance, 2.

The median frequency of topic consideration was computed for the 32 corporations on each of the 21 topics. Seven hypotheses were formulated as relevant to explanation of topic behavior variance. Objective data were obtained in order to make at least crude tests of these seven hypotheses. These hypotheses pertain to metropolitan versus non-metropolitan locus of firm, number of personnel in the firm, number of members on corporate board, per cent of board member overlap with operating management, kind of industry (most to least heavy), mean age of board members, and extent of service of average board member on other boards.

## Results

**Activity profile.** As indicated in Figure 1, the typical corporate board in this study gives relatively frequent attention through board meetings to: future business prospects (4.3 sessions per year) competition (3.9 sessions); quantity of output (3.8); distribution (3.0); and, the business cycle (2.8). Relatively infrequent topics of board attention include: voting bonuses (1.1 sessions per year) obtaining capital (1.3); relations with government (1.4); company morale (1.4); advertising (1.4); salesmanship (1.5); evaluation of key personnel (1.5); public relations (1.5); and salaries and wages (1.7). Intermediate amounts of attention are given to: labor relations (2.5); taxes (2.5); pricing (2.2); stock inventory (2.0); quality of output (2.0); relations with stockholders (2.0); and distribution of profits (2.0).

**Related hypotheses.** The seven hypotheses of meaningful relation to board behavior as previously stated find some confirmation in Table 1. Metropolitan location of a firm is associated with certain board behaviors, particularly treatment of relations with stock-

holders, voting bonuses, distribution of profits, morale, and quantity of output.

Size of the organization in number of personnel is also associated with board behavior. Boards of larger firms give greater attention to future business prospects, taxes, output, distribution, pricing, evaluation of key personnel, stock inventory, and distribution of profits.

Size of the board itself is significantly related with board emphasis on such topics as the business cycle, distribution of profits, and advertising.

Extent of board personnel overlap with operating management personnel is associated with assignment of little attention to advertising and pricing.

The heavier the industry, the less attention does the board tend to devote to voting bonuses and to quantity of output. An interesting tendency also exists for the heavy industry boards to give more frequent attention to labor relations.

Mean age of board members per board is unrelated to the board behaviors investigated in this study.

The extent to which the board is composed of members with memberships on other boards is associated with greater board emphasis upon distribution, quantity of output, quality of output, relations with stockholders, distribution of profits, advertising, the business cycle, taxes, competition, and company morale.

## Conclusions

Insofar as these data are valid estimates of activity emphases of corporate boards, the following conclusions may be warranted:

1. The frequent topics of board attention are future business prospects, competition, quantity of output, distribution, and the business cycle.
2. Moderate board attention is given to labor relations, taxes, pricing, inventory, quality, stockholder relations, and distribution of profits.
3. Relatively infrequent attention is assigned to salaries and wages, public relations, evaluation of key personnel, salesmanship,

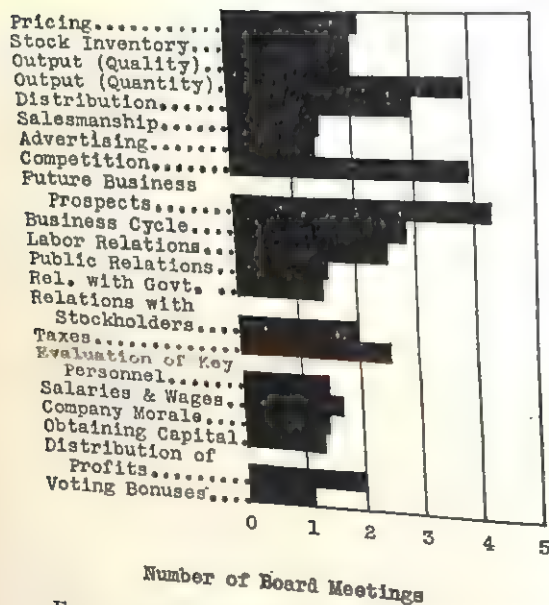


FIG. 1. Median number of corporate board meetings per year at which each of twenty-one topics is considered, according to the reports of directors of thirty-two corporations.

Table 1

Tetrachoric Coefficients\* of Correlation Between Corporate Board Emphases on Certain Topics and Seven Referent Variables

Topic	1. Metro- politan	2. No. of Per- sonnel	3. No. on Board	4. Board Overlap with Mgt.	5. Heavy Industry	6. Mean Age of Board Members	7. Responsi- bilities to Other Boards
1. Pricing	.31	.59	-.01	.49	-.17	.00	.38
2. Stock inventory	.32	.53	-.20	.04	.08	-.10	.16
3. Output (Quality)	.30	.40	.00	-.24	-.40	-.10	.61
4. Output (Quantity)	.48	.65	.19	-.35	-.60	-.21	.71
5. Distribution	.34	.60	.10	-.43	-.20	.31	.76
6. Salesmanship	-.11	.40	.18	-.43	.01	.19	.39
7. Advertising	-.08	.23	.47	-.82	-.29	.21	.61
8. Competition	.05	.41	.11	-.25	.06	.08	.50
9. Future business prospects	.09	.70	.40	-.34	.02	-.01	.05
10. Business cycle	.29	.46	.60	-.40	-.05	.07	.55
11. Labor relations	.41	.14	.10	.43	.43	-.22	-.09
12. Public relations	.41	.34	-.09	-.16	-.10	.10	.39
13. Relations with government	.16	.39	.09	-.16	.13	.08	.25
14. Relations with stockholders	.68	.40	.18	-.35	.02	-.01	.60
15. Taxes	.16	.63	.09	-.16	.13	-.10	.51
16. Evaluation of key personnel	.08	.58	.40	-.25	.19	-.26	-.09
17. Salaries and wages	.23	.28	-.01	-.28	.02	-.10	.30
18. Company morale	.50	.29	-.28	-.34	.18	.33	.49
19. Obtaining capital	-.07	.42	.31	-.16	.25	-.17	.24
20. Distribution of profits	.54	.52	.57	-.44	-.44	.08	.60
21. Voting bonuses	.67	.25	.06	.07	-.69	.23	.11

\* All coefficients for which the probability of non-chance meaning is 95 or better are indicated in italics.

advertising, morale, government relations, obtaining capital, and voting bonuses.

4. In general the topics given most frequent attention by boards are those related to immediate corporate survival, while those less frequently treated topics tend to be related either to the internal workings of the company or to special staff or usually delegated functions.

5. Such mentally stimulative factors as metropolitan environment, large number of personnel, large board, and particularly many board members who serve simultaneously on other boards are associated with more frequent consideration of practically all of the 21 topics.

6. Board overlap with operating management tends to be inversely related with frequency of consideration of the various topics. The only notable exception (significant at

non-chance probability of 90) to this generalization is frequency of attention to *labor relations*, which is considered more frequently in the "overlap" boards. This latter tendency may be due in part to defensive attitudes (defensive of management) of board members who also are a part of operating management. Insofar as this restriction of problem consideration is a result of board-management overlap, it may be a psycho-economic argument against allowing board members also to serve in operating management. These data do suggest that such overlap, when excessive, may interfere with the problem-raising and problem-solving processes in corporate enterprise.

7. Average member "responsibilities on other boards" is a variable which probably connotes experience and exceptional ability. It seems significant that boards so favored

place notably greater board meeting emphasis on competition and quality of output. It also seems of importance that none of the other six "hypothesis" variables correlates significantly with board emphasis on either competition or quality of output.

8. Mean age of board members was not a significant predictor of topics emphasized at board meetings.

*Received June 20, 1952.*

### References

1. Flanagan, J. C., et al. *Critical requirements for research personnel: a study of observed behaviors of personnel in research laboratories*. Pittsburgh: American Institute for Research, March, 1949.
2. Gordon, Thomas. *The development of a method of evaluating flying skill based on an analysis of the critical requirements of the airline pilot's job*. Unpublished Ph.D. dissertation, University of Chicago, June, 1949.
3. Wagner, R. F. Critical requirements for dentists. *J. appl. Psychol.*, 1950, 34, 190-192.

## The Curve of Output as a Criterion of Boredom \*

Patricia Cain Smith

*Cornell University*

The purpose of this study was to investigate the relationship between the experience of boredom and changes in rate of output or shape of production curves for industrial workers. The classic investigations of the British Industrial Fatigue Research Board (5, 6, 7, 8, 9) have satisfied the writers of our textbooks that the experience of monotony or boredom is characteristically accompanied by changes in the rate of output, and even that the nature of the worker's experience may be identified by examination of the shape of the curve of output. A re-examination of the work of the British investigators was made necessary by certain deviations from normally acceptable methods of scientific investigation, which will be discussed later in this paper.

As early as 1941, Roethlisberger and Dickson failed to duplicate the English results. They stated: "With respect to the monotony hypothesis, no definite conclusion could be drawn. A curve resembling what is claimed to be a typical monotony curve was not encountered except in the case of Operator 1A. It was clearly understood, however, that monotony in work is primarily a state of mind and cannot be assessed on the basis of output alone" (2, p. 127).

In 1946, Rothe undertook an investigation of the characteristics of production data, recognizing their importance as criteria in a wide variety of industrial investigations. He found that individual daily work curves "may take any of many different forms and do not assume any characteristic, predictable pattern" (3, p. 209). Correlations of work curves for the same operators for different days varied widely, the median correlation being approximately .05. Rothe averaged

work curves for each worker for one week, and obtained trend lines which were classified by inspection. Four of these curves were "mixed curves," two were "fatigue curves" and two were "monotony curves."

Rothe was interested in determining whether knowledge of the production curve for any individual or group for a specific work period would permit prediction of the characteristics of future work curves. Neither he nor Roethlisberger and Dickson attempted to relate the shape of the work curves which they obtained to the experience of the individual worker. Rothe's study, moreover, was performed using hourly-paid workers whose work flowed in a continuous and uninterrupted manner, so that his results could not be directly applied to the very different incentive conditions obtaining for piece-rate workers whose work is grouped into lots or bundles.

Since the existence of any convenient overt indicator of the psychological state of the worker would be of obvious practical importance, and would be highly useful for research purposes as well, the present investigation of the relationship between reported boredom and changes in the curve of output was undertaken. Also included in this investigation were such other proposed behavioral indices of boredom as talking, variability of production, and frequency of voluntary rest pauses.

This study was conducted in a small knitwear mill in northern Pennsylvania. Most operators in the mill, and all operators studied in detail here, were paid by piece rate. Two operations were chosen for observation. Both were: (1) short enough so that variations in production would show up in the output curves; (2) long enough to permit timing of several operators at once; (3) performed in a uniform manner by several experienced operators; and (4) largely manual, so that the operator rather than the machine de-

\* This paper is a portion of a dissertation presented as partial fulfillment of the requirements of the degree of doctor of philosophy at Cornell University. The writer is deeply indebted to Dr. T. A. Ryan for his guidance, and to the management, union officials, and workers whose active cooperation and assistance made the study possible.

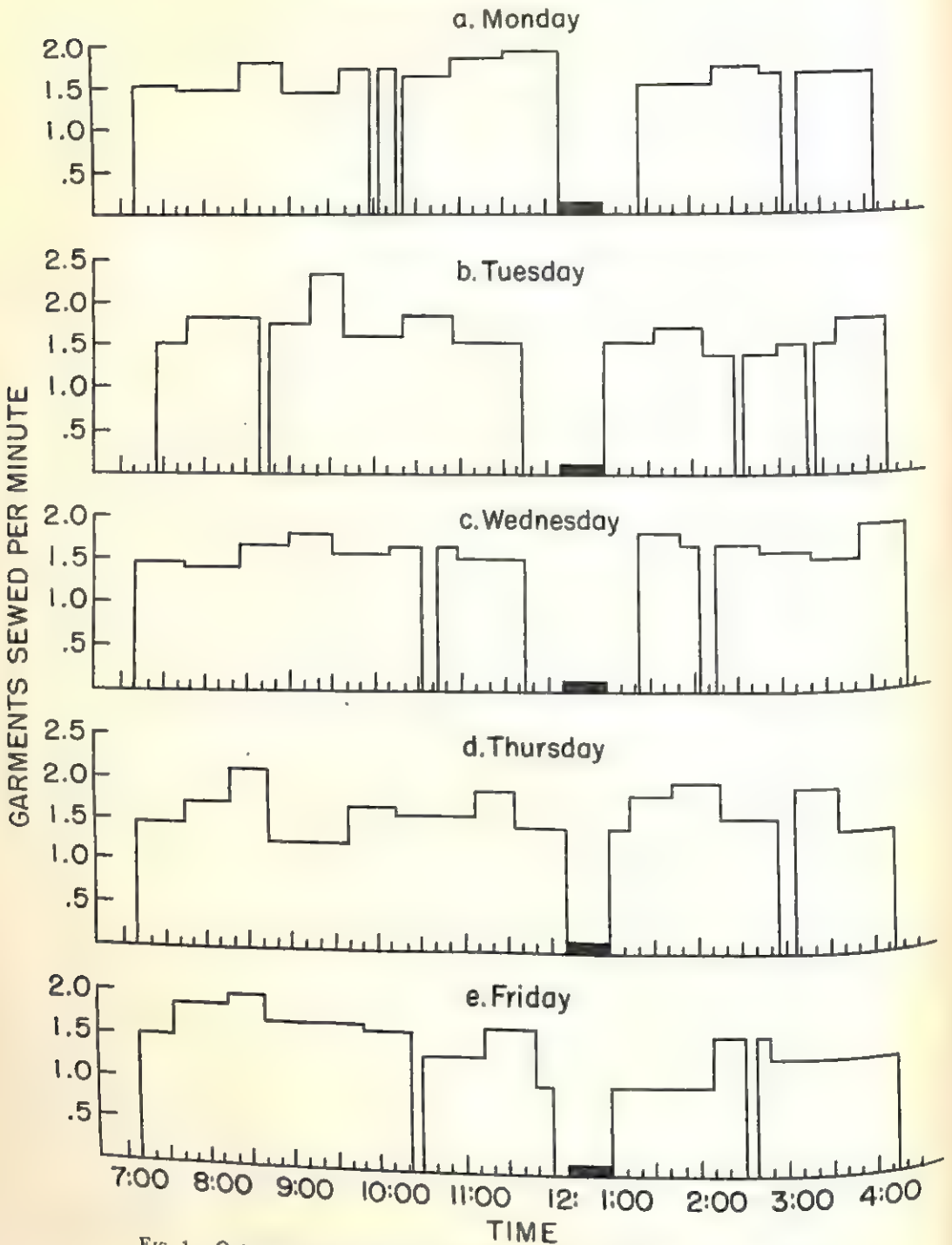


FIG. 1. Output curve for one week—Operator 7A: Hemming Operation.

judges examined the work curves for each work period and attempted to group together curves of similar shape, disregarding traditional classification systems. Groupings by the judges did not agree to a significant extent. Moreover, the groupings of neither judge agreed with either work or recreational

social groupings of the workers. None of these relationships was significantly different from chance when tested by the chi-squared test. Neither in the work nor in the recreational groups was there any evidence that work curves of members resembled one another for the same period of work.

It seemed likely that another operation would yield more traditional results. One was chosen, therefore, in which there were two portions to the task, which could be observed and timed separately. The job was called taping. Two short stiffened pieces of cotton tape were sewed on the unhemmed bottom of the shirt. After the operator finished sewing the bundle, she cut the threads and folded the shirts. Again the operators were timed continuously for one week, and again no characteristic curves were found for cutting, for sewing, or for the two combined. Similarly, there seemed to be no relationship between any daily work curve and the reported feelings of the operator. Again, the only operator who produced either ascending or U-shaped curves reported that, despite the numerous disadvantages of her work and the company, she was certainly not bored. Thus the production curve criteria proved not only unreliable, in that observers could not agree upon classification of curves, but invalid as well.

One of the major difficulties in the use of production curves as criteria lies in the analysis of data. There are no satisfactory statistical measures available for the comparison of the shapes of curves. Interpretation of the results of correlational analyses is sometimes made difficult by the peculiarities of the distributions involved. Moreover, the correlation coefficient cannot allow for over-all similarities in the shapes of the curves, when the changes of slope are displaced slightly in time. The alternative method of visual inspection is subjective and, apparently, unreliable. Both methods show little agreement from day to day, so that even though it could be demonstrated that daily curves reflected the experience of the individual workers, it would be highly unlikely that any long-term relationships could be demonstrated. Their use in this kind of situation appeared to be impractical.

Other changes in behavior which have been related to boredom include frequency of talking, frequency of rest pauses, variability in rate of working, and average speed of working. It was possible to rank the workers within each group on each of these factors

and on intensity of boredom symptoms, estimated from both questionnaire scores and interview responses. The rankings were compared and the relationships tested for significance by Kendall's non-parametric tau test (1, 403-408). No significant or even consistent relationship appeared between the boredom symptoms and the proposed indices. Reliability of the behavioral indices was estimated by comparing total rankings for each worker on Monday, Tuesday and Wednesday with the totals for Thursday and Friday. All of these relationships proved significant at the 5 per cent level or better by Kendall's tau test. Individual differences were, therefore, reasonably stable throughout the week.

### Discussion

Why were these results so different from those of the British Industrial Research Board? In the first place, comments of the workers showed that each had her own concept of the number of bundles that she should complete in a day. If she was behind schedule, she hurried toward the end of the day; if she was ahead, she slackened speed or stopped entirely. One operator, who had just completed all but one of her customary bundles for the day, commented, "You've seen how fast we can do them. Now do you want to see how slow?" Production figures reflected quite clearly what the workers considered to be the proper pace for them at that particular time, but not at all necessarily the way they felt about their work.

It has been the observation of the writer that such pacing of work occurs with much greater frequency in industrial situations than does spontaneous variation in rate. Even when there is no restriction due to fear of rate-cutting, it is normal for any worker to decide in advance how much he will produce, and earn, each day. Effort is unquestionably pegged, at least within narrow ranges, in most industrial situations.

A careful re-examination of the English studies suggests several differences in method which perhaps further account for the discrepancy between our results and theirs. The most serious has already been mentioned; they included in their criterion items which

were related to changes of rate of working, and weighted these items in the direction favorable to their hypothesis. The reader is not told, moreover, whether or not their curves were classified without knowledge of the accompanying verbal reports. Several other factors apparently operated to make the shape of their curves more consistent from day to day. Although they do not specify the kinds of jobs involved, one would infer from comparison of the various reports that at least six different operations were involved, with various hours of work and methods of payment. Such variations in jobs and conditions would tend to mask individual variability.

One last factor should be noted. There is no indication in any of their data of voluntary rest pauses, even for rest-room visits. If decreases in production due to such work stoppages were averaged into their curves, this procedure would account for the consistency of the curves from day to day, as well as for the preponderance of U-shaped curves, since rest-room and water fountain visits tend to be made at about the same time every day, and mostly in the middle of the work period.

### Summary

Continuous observation of two groups of eight women each, operating power sewing machines on light, uniform and repetitious work, led to the following conclusions:

1. There were fairly stable individual differences in speed of working, variability of production, frequency of rest pauses, and frequency of talking.
2. These differences showed no consistent relationship to the reports of the workers concerning their feelings of boredom or monotony.
3. No shape of work curve was found which would characterize the individual worker.
4. Work curves for individuals forming social groups showed no observable relationship with each other.
5. The approach of the closing hour had a noticeable effect on the production of many of the workers. The direction of the change

in rate which appeared at the end of the day was determined by the concept of a day's work held by the worker.

6. Boredom is not necessarily accompanied by a depression in the curve of output, nor is a sag necessarily accompanied by feeling of boredom.

7. Output curves should be viewed with caution as indications of the subjective feelings of the worker.

There can be little quarrel with the claim of the British investigators that, other factors being equal, workers tend to slow down, talk, become restless and variable in their production when bored. In most industrial situations, however, one cannot assume that all other factors are equal, and many of these factors may heavily outweigh the influence of interest or boredom in producing changes in working behavior.

Received May 28, 1952.

### References

1. Kendall, N. G. *The advanced theory of statistics*. London: Charles Griffin and Company, 1945, Vol. 1.
2. Roethlisberger, F. J., and Dickson, W. J. *Management and the worker*. Cambridge: Harvard University Press, 1941.
3. Rothe, H. F. Output rates among butter wrappers: I. Work curves and their stability. *J. appl. Psychol.*, 1946, 30, 199-211.
4. Rothe, H. F. Output rates among butter wrappers: II. Frequency distributions and an hypothesis regarding the "restriction of output." *J. appl. Psychol.*, 1946, 30, 320-327.
5. Vernon, H. M., Wyatt, S., and Ogden, A. D. On the extent and effects of variety in repetitive work. *Industrial Fatigue Research Board Report*. No. 26, 1924.
6. Wyatt, S., assisted by Fraser, J. A. Studies in repetitive work with special reference to rest pauses. *Industrial Fatigue Research Board Report*. No. 32, 1925.
7. Wyatt, S., and Fraser, J. A., assisted by Stock, F. G. L. The comparative effects of variety and uniformity in work. *Industrial Fatigue Research Board Report*. No. 52, 1928.
8. Wyatt, S., and Fraser, J. A., assisted by Stock, F. G. L. The effects of monotony in work. *Industrial Fatigue Research Board Report*. No. 56, 1929.
9. Wyatt, S., and Langdon, J. N., assisted by Stock, F. G. L. Fatigue and boredom in repetitive work. *Industrial Health Research Board Report*. No. 77, 1937.

## Predicting Success in Elementary Accounting

O. R. Hendrix

Office of Student Personnel and Guidance, University of Wyoming

A recent article by Traxler<sup>1</sup> regarding the use of objective tests for the selection of personnel in the professional field of accounting, encourages further investigation of the suitability of a number of tools for the prediction of success in college courses in accounting. This study represents a preliminary investigation of the relative validity of Form C of the American Institute of Accountants *Orientation Test* (AIA), the 1947 Edition of the American Council on Education *Psychological Examination* (ACE), Form 23 of the Ohio State University *Psychological Test* (OSU), and the accountant scale of the *Strong Vocational Interest Blank for Men* (SVIB), for predicting success in elementary accounting at the University of Wyoming.

In the fall of 1949, the AIA *Orientation Test* (Form C) and the *Strong Vocational Interest Blank for Men* were administered to 110 freshmen students enrolling in elementary accounting in the College of Commerce and Industry at the University of Wyoming. Scores on the other two tests mentioned above were already available for most of these students and it was possible to secure the four test scores and accounting grades for 95 students out of the 110. Of this number 76 were men and 19 were women.

Statistical constants for the four tests and the criterion are presented in Table 1.

Intercorrelations for the five variables involved were computed and are contained in Table 2.

The highest coefficient of correlation, .84, was that between OSU and ACE. This was to be expected since both are tests of general ability to do college work. A substantial relationship is also noticed between these two tests and AIA *Orientation Test*. No substantial relationship seems to exist between SVIB and the other three tests although the

coefficient of correlation of .18 between SVIB and AIA *Orientation Test* is of interest. The standard error of .18 was .099, indicating significance between the .05 and .10 levels. The most interesting revelation is that both ACE and OSU seem to be more closely related to grades in elementary accounting than is AIA *Orientation Test*. This is especially interesting in view of the fact that AIA *Orientation Test* is intended to be "a general intelligence test slanted toward business."<sup>2</sup>

Multiple correlations were computed between all possible pairings of the four tests and accounting grades. Table 3 lists the correlations obtained. If the prediction tools are to be limited to two out of the four tests considered here, it would seem that the best two combinations would be ACE and SVIB, or OSU and SVIB. Again the interesting revelation is that AIA *Orientation Test* is not to be found in either of the best two combinations of two out of four tests.

Multiple correlations were also computed between all combinations of three out of four tests and accounting grades. These correlations are recorded in Table 4.

The best combination of three tests for predicting success in elementary accounting is apparently ACE, OSU and SVIB (Account-

Table 1

Means and Standard Deviations of Test Scores and Fall Quarter Grades in Elementary Accounting\*

Variable	Mean	S.D.
ACE Psychol. Exam.	110.5	18.3
OSU Psychol. Test	67.9	24.3
AIA Orientation Test	33.9	11.0
Strong Interest Blank, Acctg. Key	36.9	11.0
Elem. Accounting Gr.†	2.9	1.0

\* N = 95.

† Grades given at the University of Wyoming are as follows: I (A), II (B), III (C), IV (D), and V (Failure).

<sup>2</sup> *Ibid.*, p. 428.

<sup>1</sup> Traxler, A. E. Objective testing in the field of accounting. *Educ. psychol. Measmt.*, 1951, 11, 427-439.

Table 2

Intercorrelations of Test Scores and Grades in Elementary Accounting for the Fall of 1949\*

	OSU Psychol. Test	AIA Orient. Test	Strong Interest Blank	Elem. Accounting Grades†
ACE Psychol. Exam.	.84	.66	.00	.36
OSU Psychol. Test		.52	-.10	.37
AIA Orient. Test			.18	.32
Strong Interest Blank, Acctg. Key				.26

\*  $N = 95$ .

† Due to the grading scheme employed, e.g., A = 1, B = 2, etc., these coefficients of correlation as compiled were all negative, but are listed here as positive since the true sense of the relationship is positive.

ing Key). Again it is interesting to note that this is the one possible combination of three out of the four tests that does *not* include the AIA Orient. Test. Addition of AIA Orient. Test to the cluster of three tests did not appreciably increase the predictive value of the cluster. (Both  $R$ 's were .55 when rounded to two decimals. Theoretically the introduction of an additional variable into a cluster will always increase  $R$ , but the increment in this instance was so small that it is not observable when the  $R$ 's were rounded.)

The findings are all based upon the as-

Table 3

Coefficients of Multiple Correlation Between Various Pairings of Test Scores and Grades in Elementary Accounting in the Fall of 1949\*

Pairs of Test Scores	$R$
ACE Psychol. Exam. and Strong Interest Blank, Acctg. Key	.51
OSU Psychol. Test and Strong Interest Blank, Acctg. Key	.48
ACE Psychol. Exam. and AIA Orient. Test	.41
OSU Psychol. Test and ACE Psychol. Exam.	.39
OSU Psychol. Test and AIA Orient. Test	.39
AIA Orient. Test and Strong Interest Blank, Acctg. Key	.39

\*  $N = 95$ .

sumption that accounting grades are an acceptable criterion for judging the relative validity of the tests under consideration. While grades are known to be not as reliable as is desired, they are the criterion of performance most generally used in college courses.

It is obvious that this study is restricted to the relationship between the test scores considered and grades. It does not necessarily follow that the same relationship exists between the test scores and success in actual employment in the field of accounting.

For example, it is possible that many college professors weigh mastery of the theoretical aspects of accounting more heavily than practical skills in accounting when awarding grades. On the other hand, success in employment in accounting may be more closely related to the practical skills. These, of course, are hypothetical assumptions, but they do illustrate the danger of drawing conclusions from this study concerning the rela-

Table 4

Coefficients of Multiple Correlation Between All Possible Combinations of Three and Four out of Four Tests and Grades in Elementary Accounting

Combinations of Test Scores	$R$
ACE, OSU, and SVIB	.55
ACE, OSU, SVIB, and AIA Orient. Test	.55
ACE, SVIB, and AIA Orient. Test	.51
OSU, SVIB, and AIA Orient. Test	.49
ACE, OSU, and AIA Orient. Test	.39

tionship between *AIA Orientation Test* and success in employment in accounting.

Summary

1. If a single test is to be utilized in predicting grades in elementary accounting, *ACE Psychol. Exam.* and *OSU Psychol. Test* are preferable to the *AIA Orientation Test*.
2. If two tests are to be used, neither of the two best combinations of two out of four tests includes the *AIA Orientation Test*.

3. If three out of the four tests are to be used, the best combination of three does not include the *AIA Orientation Test*. The addition of *AIA Orientation Test* to the cluster of three does not improve the predictive value of the cluster.

4. It does not necessarily follow that the same relationship would be obtained if the criterion used were success in professional employment as an accountant.

*Received May 28, 1952.*

# An Index of Selective Efficiency (S) for Evaluating a Selection Plan

William Leroy Jenkins

Lehigh University

Suppose a selection plan has been validated and the multiple *R* turns out to be about .60. What does a validity coefficient of this size indicate about the selective efficiency of the plan?

The index of predictive efficiency (*E*) is not a satisfactory measure. For an *R* of .60:

$$E = 100 (1 - \sqrt{1 - r^2}) = 20\%$$

which represents the per cent improvement over chance in predicting individual criterion scores. But ordinarily a selection plan is designed merely to pick a group of successful workers and to eliminate a group of unsuccessful workers—not to predict the criterion score of each individual.

What we need is an index of selective efficiency that will indicate how well we can pick such groups. Particularly we are interested in accepting as many as possible of the potentially superior workers and rejecting as many as possible of the potentially inferior workers. Let us call the highest quarter on the job criterion "superior workers," and the lowest quarter on the job criterion "inferior workers." The middle half will be "mediocre workers." Then:

Successes of the plan are superior workers accepted and inferior workers rejected.

Failures of the plan are superior workers rejected and inferior workers accepted.

Suppose we have two hundred applicants and choose half of them with the aid of a brown Stetson hat. If we obtain job criterion scores for all of them, we can expect to find something like this:

	Inferior	Mediocre	Superior
Accepted	25	50	25
Rejected	25	50	25

Successes: 25 + 25 = 50

Failures: 25 + 25 = 50

Selection on a chance basis leads in the long run to an equal number of successes and failures.

But suppose we use a selection plan having a validity coefficient of about .60. If we obtain job criterion scores for all two hundred men, we should find something like this:

	Inferior	Mediocre	Superior
Accepted	10	50	40
Rejected	40	50	10

Successes: 40 + 40 = 80

Failures: 10 + 10 = 20

Improvement over chance:  $\frac{80 - 20}{80 + 20} = 60\%$

With any actual sample of 200 applicants, the figures might not come out in this exact symmetrical pattern but the per cent improvement over chance should be substantially the same.

With the aid of a chart for computing tetrachoric *r*<sup>1</sup> it is possible to determine the theoretical improvement over chance corresponding to any obtained value of *R* for any proportion of total applicants accepted. Some typical values of the index of selective efficiency (*S*) are shown below:

Proportion Accepted	<i>R</i> = .50	<i>R</i> = .60	<i>R</i> = .70	<i>R</i> = .80
One-third	48%	57%	66%	76%
One-half	52%	63%	74%	85%
Two-thirds	48%	57%	66%	76%

For all practical purposes we may say: *the index of selective efficiency (S) has the same numerical value as the validity coefficient, if we are accepting something between one-third and two-thirds of the applicants.*

In our experience, the index of selective efficiency (*S*) has proved a useful way of explaining the meaning of a validity coefficient to someone who is unfamiliar with statistics.

Received June 16, 1952.

<sup>1</sup> Jenkins, W. L. A single chart for tetrachoric *r*. *Educ. psychol. Measmt.*, 1950, 10, 142-144.

## A Note on Techniques in the Investigation of Accident Prone Behavior \*

Lawrence L. LeShan and Jim B. Brame

Roosevelt College

University of Houston

In the past several years there have appeared, in the psychological literature, a large number of studies of accident proneness. Many of the articles which have appeared have lost some of their potential value due to a lack of clarity concerning the special problems of technique which exist in this field. It is the purpose of this paper to point up a few of these problems.<sup>1</sup>

### Method

The usual method of finding a population of accident prones includes either an interview technique or a survey of accident records in an industrial organization, a police file, insurance records or some source of this sort. Each of these has dangers attached to it.

*The interview.* Accident prones have a strong tendency to "forget" accidents. A few examples may serve to illustrate this.

One man revealed half a dozen major accidents. Intensive interview probing found no others. At the end of the interview, he was asked to strip, and his body was examined for scars. A previously undisclosed scar on the right side of his chest was called to his attention. He then remembered that three years previously a bulldozer had rolled over him, injuring his back and breaking three ribs.

Another man was leaving the interview

\* The authors accept full responsibility for this paper. It is a pleasure, however, to thank Thomas Fansler, Research Director of the National Safety Council, for raising and clarifying many of these points.

<sup>1</sup> The authors became interested in this problem as a result of being involved in research concerning the psychodynamics of individuals with a history of repeated accidents. (The results of this study were published in *Psychiatry: Journal for the Study of Interpersonal Processes*, Vol. 15, No. 1, 1952, pp. 73-80. As part of this research, thirty-five accident-prones were interviewed by one of the authors (JBB). The other part of the study (consisting of analyses of projective tests on sixty-five accident-prones and seventy-five equated non-accident-prones and approximately twenty intensive interviews with accident-prones) was completed by the other author of this paper.

room when the examiner (JBB) noticed he had a bent distal phalanx of the right little finger. On inquiry the patient said he "just remembered" that he broke that finger the previous year.

A twenty-one year old male with several accidents denied any further accidents during thirty minutes of detailed questioning. Towards the end of this period he started rubbing his right elbow. On specific questioning, he recalled that he had broken his arm when he was eighteen.

Behavior of this sort is by no means infrequent. In the experience of the authors, it is the general rule rather than the exception.<sup>2</sup>

For this reason, a great deal of skepticism must be attached to results gained by the written questionnaire method also. An experience of one of the writers (LLL) illustrates this point. A questionnaire was filled out by 40 accident repeaters who had been called into a state driving clinic. They had all had at least three auto accidents in 12 months. Over half of them did not remember all three accidents.

When an interview technique is being used to obtain an accident history, the subject should be questioned on a year-by-year basis. This would include the jobs worked at and the particular hazards of each job; vehicles driven, repairs and their cost; sports participated in, falls and bruises. Special reference should be made to burns and scalds since these are not often thought of as "accidents" by the subject.

The interview frequently makes people defensive about their accidents record as they

<sup>2</sup> No quantitative estimate as to how large a percentage of their accidents these individuals forget can be made, since we do not know how many accidents were not recalled at all in our interviews. However in thirty-five interviews, at least thirty of the subjects recalled several more accidents after careful probing than they had when simply asked to list all the accidents that they had had and then were given plenty of time and a sympathetic listener.

may see implications of punitive intent. They may, therefore (in addition to the accidents that they have repressed), deliberately not state others which they do consciously remember. For this reason, careful attention must be paid to the psychological atmosphere of the interview. A good relationship is essential to accurate data collation. We feel that, by and large, an authoritarian relationship tends to produce markedly less data than an egalitarian one.<sup>3</sup> A procedure that is often helpful is to express interest in the general health history and to record all illnesses.

One point about the interview which should be considered in research design is that it is essential to gather data on control groups in the same manner it is gathered on groups of accident repeaters. An intensive probing interview covering the entire life-span of the individual produces a surprisingly large number of accidents in the general run of the population. Since a definition of accident proneness implies that the individual concerned has a higher accident rate than his peers, both experimental and control groups have to be evaluated with the same technique.

*The use of accident records.* Probably the most common technique for studying accident-prones is to use the data of the safety departments of police or insurance firms, industrial firms, etc. There are several dangers inherent in this method, two of which might be mentioned briefly.

Although this is probably a valid way of collecting data on experimental groups, it is a dubious procedure for control groups. We do not know how many individuals are accident prone at home and not at work. If a man has a high off-duty accident record and a low on-duty accident record and we study him as a non-accident prone since we have only the plant statistics, he is likely to confuse our data, to say the least.

We know so little about the accident-prone that we do not know if he is more or less prone to report his accidents to the plant infirmary or to the police, if he tends to report only certain types of accidents, etc.

<sup>3</sup> This statement is not the result of any experimental work, but simply an impression based on experience with varied types of interviews.

## Defining an Accident

This is a complex and difficult problem. Generally we consider an accident to be a mishap with a sudden onset. However, this by no means solves our problems. Parenthetically, it might be stated that the Workmen's Compensation Act of the State of Virginia has a four line definition of "injury" which is followed by seventeen single spaced pages of clarification in fine print.

We have little clear understanding of the difference between a disease and an accident. If a workman habitually neglects washing his hands after he finishes work with coal-tar products and develops a skin irritation which incapacitates him, how does this differ from typical accident-prone behavior in which the individual injures himself by neglecting elementary safety precautions? Should we count this as an accident?

Even though we eliminate occupational disease and use only traumatic injury, other problems arise in the same area. We see a report of a man who has 15 back-sprains. The medical report states he has a "weak back." Is each sprain to be counted as an accident? Is there a difference between the man who has this particular disorder and (granted freedom of choice) repeatedly gravitates to jobs calling for heavy lifting and a similarly handicapped man who takes positions which will not put such a strain on his back?

The difference between a chargeable and a non-chargeable accident is often used in studies but is frequently more apparent than real. Surveys of trucking company records by one of the writers (LLL) have shown that individuals who have high rates of chargeable accidents tend also to have high rates of non-chargeable accidents. Many accidents which appear to be non-chargeable on superficial examination are chargeable when carefully examined. One accident prone had had four automobile accidents while he was sitting in the front seat of a car and someone else was driving. He had, he said, "generally been talking to the driver when it happened." In one of the accidents he had hurt his elbow badly as it had been outside the ventilator window when the car crashed. This state-

ment stimulated the interviewer to probe at some length into exactly what had happened. After five minutes a picture emerged that was quite different from the earlier "non-chargeable" one. It is true that he had been sitting beside the driver, but he had decided to clean the windshield. He thrust his hand with a towel through the ventilator window. At 50 MPH, the towel flapped over and covered most of the windshield, the driver was blinded and the crash occurred.

Another type of problem in defining the accident is illustrated by an individual who had no history of injuries or accidents (as they are usually defined). However, investigation showed that he had been fired from his last position (as a pharmacist) for "making mistakes." At the time he was seen he was working as a pilots' mechanic. This man had no automobile crashes or falls in his background. He simply made minor errors in work of such a nature that the errors could have disastrous effects. Definitions of accident made for a particular study should clearly exclude or include individuals of this sort.

### General Considerations

There are no agreed-on definitions of "accident," "injury," or "accident prone." Each study must first decide what it is attempting to find out. In terms of various factors such as population studied, purpose of research, techniques available, etc., definitions can be made.

This, perhaps, can be most clearly seen in defining the accident-prone. There is general agreement that he should have an accident rate higher than that of his peers, but as to how far above the mean of his peer group he must be there is no agreement. Shall we cut off the upper 1% of our population and label them "accident-prones," or shall we use the upper 5%, the upper 25%, or the upper 50%? There is no agreement here.

The problem can be approached in another way. Rather than examine (by implication) the accident liability of the specific environment we are studying, as was implied in the last paragraph, we can examine the accident liability of the individual. We can then use criteria such as one accident per

year for at least 5 years, or 3 accidents every 2 years for 10 years, etc. In this way the State of Oregon labels a man an "accident-repeater" if he has had 3 accidents in any 12-month period. This only includes 4% of state drivers, but the 4% have 40% of the accidents. (Unlisted memo. in the files of the National Safety Council.)

A problem here is that the total accident record of some individuals is not consistent by a year-by-year, or even decade-by-decade, analysis. Often a person may normally be non-accident-prone but for a period of two to five years show high accident rate and then at the end of this time, return to his former low level of accidents.

In the design of research, it may be unwise to use as controls only individuals with low accident rates. There is no evidence that this is not a special group with different characteristics than are found in the normal population. Until this problem has been investigated, controls should be taken from the center of the accident curve rather than from the lower extremity.

Good research design will demand that account be taken of both the accident liability of the specific environment and the liability of the individual. Fleming and Dickinson's excellent paper<sup>4</sup> discusses the relationship of personal and situational liability. They state, in part, "A high accident potential and an accident-prone driver make for a high accident expectancy. A high accident potential and a normal driver make for an accident possibility" (p. 171). No study which does not evaluate both the individual and the group accident rate can expect to produce clear cut results.

### Summary

Special problems exist in the design of studies of accident prone behavior. A few of these are briefly discussed. Difficulties in finding the accident-rate of an individual, defining an accident, and delimiting accident-prone and the non-accident-prone groups are pointed out.

*Received May 26, 1952.*

<sup>4</sup> J. Fleming, Jr., and J. J. Dickinson. Accident proneness and accident law. *Harv. law Rev.*, 1950, 63, 169.

## The Efficiency of the Minnesota Teacher Attitude Inventory for Predicting Interpersonal Relations in the Classroom \*

Robert Callis

*University Counseling Bureau, University of Missouri*

Our main problem in this study is to test the efficiency of a measuring instrument to predict the ability of a teacher to effect harmonious interpersonal relations in the classroom. We believe that harmonious interpersonal relations in the classroom are desirable. We also believe that the teacher is a key figure in the kind of relationship that prevails. If good interpersonal relations are obtained between teacher and students, then it follows that the teacher and students will work together in a social atmosphere of co-operative endeavor and with a mutual feeling of security. Also the students will be motivated to learn the material at hand more easily, and will have an opportunity to do so in a manner which is most efficient for them individually. If, on the other hand, the social climate in the classroom is characterized by tension, fear, and submission on the part of the students, the student is apt to have little motivation to learn; and, as a by-product, numerous disciplinary problems, inattention, and restlessness will result. If there is mutual distrust and hostility between the teacher and students, probably little learning will occur.

We have assumed that a teacher's attitudes resulting from his life experiences will have a noticeable effect on the kind of relationships which this teacher creates in his classroom. These attitudes are presumed to be a result of a multitude of factors such as values, personality traits, intelligence, general knowledge, and teaching skills. If we

are able to measure these attitudes satisfactorily, we then should be able to predict to a significant degree the kind of relationship which will be obtained in the classroom. Specifically our problem is: *How well will the Minnesota Teacher Attitude Inventory (MTAI) predict interpersonal relations in the classroom?*

### Procedure

*The predictor.* The MTAI was selected as the predictor for this study since it attempts to measure the kinds of teacher attitudes which are relevant to teacher-student relations. Two studies of the validity of the MTAI have already been reported (2, 3). In each of these it was found that the MTAI would predict a three-fold criterion of teacher-student relationships to the extent indicated by a correlation coefficient of .59. The MTAI contains 150 attitude statements to which the teacher responds with one of five possible responses. The scoring system was determined by purely empirical means (2). Responses to the MTAI were secured from one group of teachers judged to be superior in their relations with students and another group judged to be inferior in their relations with students. The per cent of each group choosing the various response categories was computed and the significance of the difference between these percentages was determined. A significant difference in percentage favoring the superior group was scored "+ 1"; a significant difference favoring the inferior group was scored "- 1"; all non-significant differences were scored "0." Following is an example:

\* This study was a part of the University of Missouri Agricultural Experiment Station Project No. G-48, which was a part of the Office of Naval Research Project NR 154-111.

Item: Most children are obedient.					
	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Superior group	34%	58%	4%	3%	1%
Inferior group	18%	64%	4%	13%	1%
Differences in %	+16	-6	0	-10	0
Scoring	+1	-1	0	-1	0

It can be argued, on logical grounds, that the "uncertain" and "strongly disagree" response categories should be scored "-1." However, in the past logical face validity for determining scoring systems has been found to be such a notoriously poor predictor of psychological functions, that the authors of the MTAI decided to use a scoring system based on empirical data only.

*The criterion.* A major task was to describe adequately the kinds of relationships which existed in each of several classrooms. We obtained three estimates of this relationship from different sources. First we obtained such an estimate from the students in each classroom. This was obtained through a 47-item questionnaire or inventory about "My Teacher," which was administered to all students in attendance the day we contacted the class. This inventory is the same as the one used by Leeds (2). Such questions as these were asked: "Do you like school?" "Is this teacher often bossy?" "Is this teacher usually kind to you?" "Is this teacher usually kind to you?" "Is this teacher usually kind to you?" The inventory was scored "rights minus wrongs." The possible range in scores was +47 to -47. Therefore, a score of zero indicates that the student made as many negative criticisms of the teacher as he made positive statements about him. The zero score would be below that expected for an average teacher. The mean score on the student inventory for each class was obtained. This mean score constituted the evaluation by the students of the interpersonal relations in that particular classroom.

The second evaluation of the classroom relations was made by the principal of the school. The principal made his evaluation in the form of a rating scale. This is the same rating scale which was designed and used by Leeds (2). Items 1 through 6 of the rating scale were scored on a 5-point scale, thus yielding a possible range in scores of 6

through 30. When the ratings were inspected, it was found that there were wide discrepancies among the means of ratings made by the various principals. We considered that these discrepancies could be due to: (a) wide variation in the leniency of the raters; or (b) wide differences in the quality of teachers in the various school buildings. It was necessary to assume one or the other in analyzing our data. We chose the former. Consequently, the principal ratings were expressed as deviations from the mean of the particular rater. That is, all the ratings which each principal made were averaged and each teacher's score was expressed as a deviation from that mean. In this way we, in effect, equated all schools on the quality being rated. This assumption of equality of all schools in classroom relations is not necessarily justified, but it was our opinion that less error would result with this technique than to assume that all raters (principals) were equally lenient in their ratings. It would have been desirable to equate the variability of each set of ratings in addition to equating the means. This was not done because several sets of ratings contained only three or four cases.

The third estimate of the classroom relations was made by two observers from our research team. Each observer visited the classroom at different times and observed the class in process for thirty minutes to an hour. Independent of each other they recorded their observations on a rating scale. Items 1 through 5 on the rating scale were scored on a 5-point scale for each item, thus yielding a possible range in scores from 5 through 25. Each observer's ratings were converted to standard scores (based on his own distribution) and then the two standard scores were averaged to arrive at the criterion of "mean observer rating." The "mean observer ratings" constituted the third

estimate of the criterion of classroom relations.

Each of the three above criteria—student ratings, principal ratings (deviation scores), and mean observer ratings—were converted to standard scores and summed. The sums of the three criteria scores were converted to standard scores and called the composite criterion. This last step was done merely to facilitate inspection of our data.

*The sample.* The sample for this study consisted of 77 public school classes in central Missouri. Grades four through ten in four school systems were represented. The population of these cities varied from 7,500 to 26,000. There was only one teacher contacted in these four school systems who declined to participate in the study. There were 82 classes in the original group from these four cities. The sample was reduced to 77 due to incomplete data. In these four cities all Negro children attended a school separate from the schools for the white children. None of the Negro classes was included in this study. The grades included varied from city to city, depending upon the organization of that particular school system. In grades four through six, the classes met as the usual elementary school class. In grades seven through ten, the grades were organized on a typical high school plan. Of the 77 teachers there were 8 male teachers, 48 married female teachers and 21 single female teachers.

MTAI, Form A,<sup>1</sup> was administered to each teacher in the study. The relationship was determined between this predictor and each of the three estimates of the criterion, and the combined criterion.

Results

Table 1 presents the means and standard deviations for the predictor and the various criteria. The means of the ratings made by our two observers are quite similar; however, the variance of ratings made by observer X was significantly greater than for observer Y ( $F = 2.13$ ;  $n_1 = n_2 = 76$ ;  $P < .01$ ).

<sup>1</sup> This was the unpublished form of the MTAI. The form which was published subsequent to this study (1) varies in a few minor details only from the one used here.

Table 1

Summary Statistics for the Predictor (MTAI) and the Various Criteria  
Note:  $N = 77$  classes.

Variable	Mean	Standard Deviation
Criteria:		
(1) Observer X Ratings	18.7	4.8
(2) Observer Y Ratings	18.1	3.3
(3) Mean Observer Ratings	49.8†	10.1†
(4) Student Ratings	24.4	11.4
(5) Principal (deviation score)	50.4†	9.3†
(6) Composite	49.7†	10.1†
Predictor:		
MTAI	27.5	35.4

† Based on standard scores computed from a few more cases than the 77 teachers in the correlational analysis. Values other than those marked (†) are based on raw scores.

The two mean ratings are only slightly greater than the middle or average point on the rating scale. This agrees with our more subjective impression that we were dealing with a typical or average group of teachers.

The ratings made by the principals are in sharp contrast with those made by our observers. The mean of ratings made by the principals was 25.0, while the highest possible rating was 30. The principals' ratings would characterize the group of teachers as highly superior in their relations with students.

The mean score on the student inventory for all the teachers was 24.4, where the possible range of score was + 47 to - 47. There are no norms available with which to compare this value.

The mean MTAI score of 27.5 is estimated to be about average or slightly below average for experienced teachers. No directly comparable norm group was available; however, norms on somewhat similar groups (beginning teachers and graduate students in education who had at least two years' teaching experience) suggest the above interpretation.

There appears to be fair agreement among the means of the various measures with the exception of the principals' ratings.

The intercorrelations among the various criterion measures, as shown in Table 2, were

Table 2

Intercorrelation of the Predictor (MTAI)  
and the Various Criteria  
Note:  $N = 77$  classes.

	Stu- dents' Ratings	Prin- cipals' Ratings	MTAI
(1) Observers' Mean Ratings <sup>1</sup>	.29**	.12	.40**
(2) Students' Ratings		.46**	.49**
(3) Principals' Ratings (deviation scores)			.19
(4) Composite of (1), (2), (3)			.46**
(5) Composite of (1), (2)			.50**

<sup>1</sup> The correlation between the ratings of the two observers was .33.

\*\* Significantly greater than zero at the 1 per cent level of confidence.

quite low with the exception of the correlation of .46 between principal and student ratings. The correlations between the MTAI scores and the various criteria, singly and combined, were significantly greater than zero except for the principals' rating: students' ratings = .49; mean observers' ratings = .40; principals' ratings = .19; composite of the three criteria = .46; and the composite of observers' and students' ratings = .50. Thus, it appears that with the MTAI we can predict the kind of interpersonal relations which will exist in the classroom about as well as we can predict academic performance by use of intelligence tests. Presumably we are measuring an aspect of personality which we

may refer to as "teaching personality." By "teaching personality" we mean those characteristics of the teacher's behavior tendencies which are associated with the teacher's ability to establish harmonious working relations with students.

The results of this study are in general similar to the ones conducted by Leeds (2, 3). The one major discrepancy between these similar studies is in the principals' ratings. In Leeds' studies the MTAI scores correlated with the principals' ratings with coefficients of .43 and .46. This is a somewhat higher coefficient than that obtained in the present study. The correlations of the MTAI scores with each of the other estimates of the criterion, that is, the observers' ratings and the students' ratings, were rather similar in the two studies. The correlation of MTAI scores with the composite criterion in this study was .46 as compared with .59 and .59 in Leeds' studies. It would appear then that we have a good start in finding predictors for our criterion of human relations in the classroom.

Received June 16, 1952.

### References

1. Cook, W. W., Leeds, C. H., and Callis, R. *The Minnesota Teacher Attitude Inventory*. New York: Psychological Corporation, 1951.
2. Leeds, C. H. A scale for measuring teacher-pupil attitudes and teacher-pupil rapport. *Psychol. Monogr.*, 1950, 64, No. 6 (Whole No. 312).
3. Leeds, C. H. A second validity study of the Minnesota Teacher Attitude Inventory. *Elem. Sch. J.*, 1952, 52, 398-405.

## Fakability of the Jurgensen Classification Inventory

H. P. Longstaff

University of Minnesota

and

C. E. Jurgensen

Minneapolis Gas Company

With the increasing use of psychological tests by industry as an aid in selecting personnel, considerable interest has developed in the problem of malingering on such tests. This is especially true of the pencil and paper type of interest and personality inventory. Bordin (1), Longstaff (4), and Strong (7) have shown that the *Strong Vocational Interest Blank* is fakable. Longstaff (4) also demonstrated the fakability of the *Kuder Preference Record*. Meehl and Hathaway (6) have found the *Minnesota Multiphasic Personality Inventory* fakable. Tiffin (8) reports a study which showed that the *Humm-Wadsworth Temperament Scale* could be faked. Wesman (9) has demonstrated similar results for the *Bernreuter Personality Inventory*.

At least four different methods have been used to try to correct this weakness. One: The development of scoring keys to detect faking. The "L" scale of the MMPI (2) is a good example. Two: Development of "suppressor variable" scales to correct scores for malingering. Notable in this connection is the work of Meehl and Hathaway (6). Three: Development of keys based on subtle and obvious items (10). Four: Development of tests using the "forced-choice technique" which it was hoped would be less susceptible to faking. An example of this approach and the subject of this paper is the *Jurgensen Classification Inventory* (3). The essential feature of such inventories is to force the subject to choose what he considers the best and worst items from a group of items which represent only "good" or only "bad" traits. It was hoped that this would get away from the weakness of presenting lists of intermixed "good" and "bad" traits where a malingerer could state that he had only "good"

traits and did not possess any of the "bad" ones.

Mais (5) developed and cross-validated a "self confidence" key for the *Jurgensen Classification Inventory*. He then had college students take the test honestly and dishonestly, i.e., trying to fake a high score in self-confidence. He found the mean score for his group changed from -5.9 (honest) to 6.9 (faked). This difference of 12.8 was significant at the .01 level. The Pearsonian correlation between the two scores was only .17.

Mais' study gave adverse data on the *Jurgensen Classification Inventory*. However, it was not a crucial study. The *Classification Inventory* was developed for use in personnel selection. In such industrial use, applicants do not know what "traits" are being measured. In fact, most keys are not based on traits but on over-all job success. It may be one thing to raise a score on a specified and named trait such as self-confidence and another thing to obtain a higher score on undefined job success.

The present study was designed to investigate further the fakability of the *Classification Inventory*. Two groups of University of Minnesota students in personnel psychology courses served as subjects. Group A consisted of 41 juniors, seniors and graduate students, the majority of whom had completed numerous courses in psychology and industrial relations. Group B consisted of 37 extension division students in an evening class, and represented a less highly selected group than the first.

### Method

Each student took the *Classification Inventory* under three sets of conditions: (1)

honest, (2) fake good over-all, and (3) fake high self-confidence. Directions for the test under these three sets of conditions were as follows:

1. *Honest score.* "This test has been constructed quite differently from most personnel tests. It has been tried out in industry and has been phenomenally successful in certain instances. Since you are students of personnel psychology, my purpose in giving you the test is threefold: *first*, I want you to become acquainted with it by actually taking it; *second*, your standing in the test may be of assistance to you in planning your vocational future; *third*, we hope to build a key that will assist us in directing students toward or away from personnel work as a vocation. Please answer the questions as accurately as you can as they apply to yourself. It will be obvious from the questions that there are no right or wrong answers. It is wholly a matter of personal preference on your part; therefore, answer the questions as they apply to you."

2. *Fake over-all good score.* "Last time, you took the *Classification Inventory* under conditions in which you were instructed to answer the ques-

tions as they apply to you. In taking the test this time imagine yourself in an employment department of a large and prosperous company. You have finished your education and are now starting out upon your life's work. You want very much to get a job with this company and hope to spend the rest of your life working for them. Therefore, you want to make as good an impression as you can. Answer the questions so you will appear in the most favorable light to the personnel manager."

3. *Fake high self-confidence score.* "You have taken this test twice before. Today I would like to have you take it trying to fake your answers so as to make a high score in self-confidence."

Means and sigmas of scores obtained under the three conditions are given in Tables 1 and 2. Results support Mais (5). Students significantly increased their scores in self-confidence when they attempted to do so, the increase averaging approximately one sigma. This increase is both statistically and practically significant. Statistical significance ( $t = 8.75$ ) is beyond the .01 level.

Table 1  
Mean Scores on Self-Confidence Key Under Three Sets of Conditions

	Honest or Accurate Score	Fake Over-all Good Score	Fake High Self- Confidence
Group A ( $N = 41$ University students)	-2.37	.07	12.22
Group B ( $N = 37$ Extension students)	2.14	2.95	10.45
Total Group ( $N = 78$ )	-.23	1.44	11.42

Table 2  
Variability of Scores (Sigmas) on Self-Confidence Key Under Three Sets of Conditions

	Honest or Accurate Score	Fake Over-all Good Score	Fake High Self- Confidence
Group A ( $N = 41$ University students)	9.52	6.71	10.62
Group B ( $N = 37$ Extension students)	11.46	11.43	11.73
Total Group ( $N = 78$ )	10.72	9.36	11.19

The situation is different when we compare "honest" scores with attempts to fake "over-all good" scores. The increase is neither statistically nor practically significant. Obviously, these students were unable to increase their scores in self-confidence when attempting to appear in the most favorable light. However, the similarity of mean scores does not give the whole story; and the test is not as satisfactory for employment use as might appear. The Pearsonian correlation between "honest" and "fake over-all good" scores was only .28. Obviously, many of the students had *attempted* to increase their score. Although scores in general were not *raised*, they were *changed*. This, of course, would be a serious defect if the test were being used for selection.

Consideration of the foregoing findings raises an important question. Did students change their answers because they thought they could improve their scores or because they were instructed to fake? *Essentially, they were directed to change answers and perhaps we should not be too surprised when they follow instructions.*

To investigate this question, another group of 68 students comparable to the previous Group A was given the *Classification Inventory* with instructions which avoided direct orders to fake answers and which simulated more nearly industrial selection and vocational guidance conditions. Directions for the test under these two conditions were as follows:

1. *Industrial selection.* "In taking this test make the following assumptions. You have just finished your college work and are in the employment department of the organization you hope to work for, applying for a job. This job you are applying for is exactly the kind of job you want so it is very important to you that you get it. The personnel manager informs you that the company has a battery of tests they give all their applicants and says, 'This is the first test in the battery. It is called the *Classification Inventory*. You will please read the directions and then answer the questions.'"

2. *Vocational guidance.* "At the last meeting of the class you took the *Classification Inventory* assuming you were applying for a job. Today I would like to have you take the test again, making the following assumptions: You are having a great deal of trouble trying to decide what voca-

tion you should go into. You finally decide to go to *The Student Counseling Bureau* to see if they can give you any assistance. The counselor informs you, 'We have a battery of tests we should like to have you take. We have found the results very helpful in dealing with problems like your own. The first test in the battery is called the *Classification Inventory*. Will you please read the directions and then answer the questions.'"

Again it was found that mean scores were essentially the same. The mean for the "Industrial" situation was  $-1.28$  (sigma of 8.98) and that for the "Vocational Guidance" situations was  $-2.18$  (sigma of 9.28). This difference is not statistically significant ( $t = .81$ ).

The correlation between "Industrial" and "Guidance" scores is .50. This is a substantial increase from the former coefficient of .28. It is apparent that the degree of faking is materially reduced by avoiding the direct suggestion to change answers. It should be pointed out that we are dealing in these experiments with a very intelligent and psychologically sophisticated group of subjects. That this is an important factor can be seen by comparing the results of Group B (Table 1 and 2) with the other groups. The extension students (Group B) increased their scores considerably less than did the other groups comprised of more highly selected students. Be this as it may, the fact clearly stands out that all three groups materially changed their answers and scores under the different sets of conditions. Although modification of directions toward greater realism decreased the extent of change, the resultant correlation of .50 is not encouraging. Obviously *faking is possible* in the *Classification Inventory*, and probably occurs when the instrument is used for employee selection purposes. Unfortunately, the extent of such faking cannot be determined for any single applicant.

Although various forced-choice tests differ in the way in which items are selected, these differences would not appear to be related to attempts to fake. Presumably, findings from these experiments might well be expected to apply to the forced-choice technique in general.

## Summary

1. Scores on self-confidence were significantly raised when students attempted to raise their scores and knew the test measured self-confidence.
2. Scores on self-confidence were not increased when students attempted to fake good over-all scores when students did not know that the test was scored for self-confidence.
3. Scores on self-confidence were not increased when students changed from a simulated "industrial" to a simulated "guidance" frame of reference when the students did not know that the test measured self-confidence.
4. The way in which instructions were worded materially affected the extent of attempted faking.
5. Although mean scores were not increased when students did not know what trait was being measured, individual scores were frequently changed to a considerable extent. This was evidenced by correlation coefficients far lower than reliability coefficients. So far as score interpretation is concerned, the attempt to improve scores is probably as important as the ability to improve scores. How should a score at the fiftieth percentile be interpreted? Does it reflect an average amount of the trait being measured? Is it the result of a successful attempt to raise a low score? Or is it the result of an unsuccessful attempt to further increase what is already a high score? The answer is unlikely to be known in any single case.
6. The *Classification Inventory* is not recommended for use in situations where persons are likely to be motivated to obtain good scores.

7. Although these data were obtained on the *Classification Inventory*, this is no reason to believe that different results would be obtained from any other forced-choice personality test.

8. It is the opinion of the authors that techniques other than the forced-choice technique will have to be devised if the problem of malingering on personality tests is to be overcome.

Received May 22, 1952.

## References

1. Bordin, E. G. A theory of vocational interests as dynamic phenomena. *Educ. psychol. Measmt.*, 1943, 3, 49-75.
2. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
3. Jurgensen, C. E. Report on the Classification Inventory, a personality test for industrial use. *J. appl. Psychol.*, 1944, 28, 445-460.
4. Longstaff, H. P. Fakability of the Strong Interest Blank and the Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 360-369.
5. Mais, R. D. Fakability of the Classification Inventory scored for self confidence. *J. appl. Psychol.*, 1951, 35, 172-174.
6. Meehl, P. E., and Hathaway, S. R. The K factor as a suppressor variable in the MMPI. *J. appl. Psychol.*, 1946, 30, 525-564.
7. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford, Calif.: Stanford University Press, 1943.
8. Tiffin, Joseph. *Industrial psychology*. New York: Prentice-Hall, 1942, pp. 117-118.
9. Wesman, A. G. Faking personality test scores in a simulated employment situation. *J. appl. Psychol.*, 1952, 36, 112-113.
10. Wiener, D. N. Subtle and obvious keys for the MMPI. *J. consult. Psychol.*, 1948, 12, 164-176.

## The Relationship Between the Judged Desirability of a Trait and the Probability That the Trait Will Be Endorsed \*

Allen L. Edwards

*The University of Washington*

There is a rather common suspicion among many psychologists that subjects tend to give what are considered to be socially desirable responses to items in personality inventories. This suspicion has been given public expression in a recent article by Gordon (3, p. 407) who comments upon "... the motivation of a majority of respondents to mark socially acceptable alternatives to items, rather than those which they believe apply to themselves."

We have here two problems. One concerns the truthfulness of a subject's answers to items in a personality inventory, i.e., whether the response accurately describes the subject. The answer to this question implies that we have available some independent criterion in terms of which the inventory response is to be evaluated. The other problem concerns the relationship between a subject's response to an item and the social desirability of that item, i.e., whether the subject tends to give a positive answer to an item that is socially desirable and a negative answer to an item that is not. The answer to this question implies that we have available some measure of the social desirability of the item to which the response can be related. It is this problem we wish to report upon here.

### The Present Study

The hypothesis to be investigated may be stated in this way: If the behavior indicated by an inventory item is socially desirable, the subject will tend to attribute it to himself; if it is undesirable, he will not. This hypothesis may be put more precisely: The probability of endorsement of personality

items is a monotonic increasing function of the scaled social desirability of the items.

To study the relationship between the probability of endorsement of personality trait items and the social desirability of the items requires that we determine independently two measures: the probability of endorsement and the social desirability scale value of the items. This study thus consists of two parts: in the first, the scale values of the items are determined; in the second, the probability of endorsement is related to the independently determined scale values.

### Determining the Scale Values

A total of 140 personality trait items, based upon Murray's (4) discussion of needs, were written and edited. The items were selected so that 14 needs were investigated with 10 items supposedly indicative of each need. The items were arranged in 10 sets of 14 items each, so that each set consisted of one item relating to each of the needs.

The items were presented to subjects with instructions to judge the degree of social desirability of the behavior indicated by each item in terms of how the behavior would be regarded in others. Judgments were made in terms of nine successive intervals, with the lowest interval representing extreme undesirability and the highest extreme desirability. The rating system was explained in terms of a sample set of four items for which judgments had already been obtained. After these ratings had been discussed, the instructions to the subjects concluded with the following statement:

"Indicate your own judgments of the desirability or undesirability of the traits which will be given to you by the examiner in the same manner. Remember that you are to judge the traits in terms of whether you consider them desirable or undesirable in others."

\* This paper was presented before the Western Psychological Association, Fresno, California, April 26, 1952. It is part of a research program made possible by an appointment as a Faculty Research Fellow of the Social Science Research Council.

Be sure to make a judgment about each trait."

The subjects judging the desirability of the items consisted of 86 men and 66 women, a total of 152 subjects. Twenty-six of the subjects were under 20 years of age, 97 were between 20 and 30 years of age, and 29 were over 30 years of age.

Cumulative distributions of the judgments were made separately by age and by sex groups. For each item we then found the interval in which the median of the distribution of judgments would fall.

In Figure 1, we show the plot of the women's intervals against the corresponding values for the men. It may be noted that in the case of only two items would the medians be separated by as much as two intervals. For 43 of the items the medians might possibly be separated by as much as one interval. For the remaining 95 items the medians would all fall within the same interval.

In the case of many of the items falling outside the principal diagonal of Figure 1, the medians would still be approximately the same for the reason that the medians of both distributions are close to the limit of the interval, but one happens to fall slightly above and the other slightly below the limit.

A similar analysis of the judgments was made in terms of the age variable. Examination of the separate distributions indicated that the scale values that would thus be obtained would be comparable and that little distortion would be introduced by pooling the judgments for all groups.

On the basis of the combined distributions, the scale values of the 140 items were found. The scale values were determined by the *method of successive intervals* (1). This method of scaling does not involve any assumption of equality of the successive rating intervals.

After determining the widths of the successive intervals and the scale values of the items on the psychological continuum of social desirability, an internal consistency test was applied (1). Using the 147 parameters calculated from the data, it was possible to reproduce the 1,120 independent, empirical

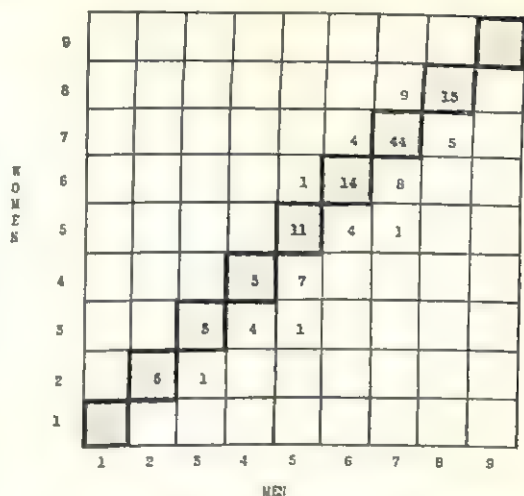


FIG. 1. Interval in which the median of the women's distribution of judgments would fall plotted against the interval in which the median of the men's distribution of judgments would fall.

observations with an average error of .023. This value, it may be mentioned, compares favorably with that usually obtained from internal consistency tests used when stimuli are scaled by the *method of paired comparisons*.

#### Relationship Between Scale Values and Probability of Endorsement

In the second part of this study, a sample of 140 pre-medical and pre-dental students responded to the same set of items for which we had previously determined the scale values on the psychological continuum of social desirability. This time, however, the items appeared in a printed form as a personality inventory. The inventory was part of a test battery which was administered for the Medical and Dental Schools of the University of Washington. The instructions were those that are commonly used with personality inventories. A "Yes" response indicated that the subject believed that a given item was characteristic of himself and a "No" response that it was not.

Item counts were made for each item, by means of IBM equipment, and the per cent responding "Yes" was then found for each item. This per cent is the proportion of the sample indicating that the behavior stated

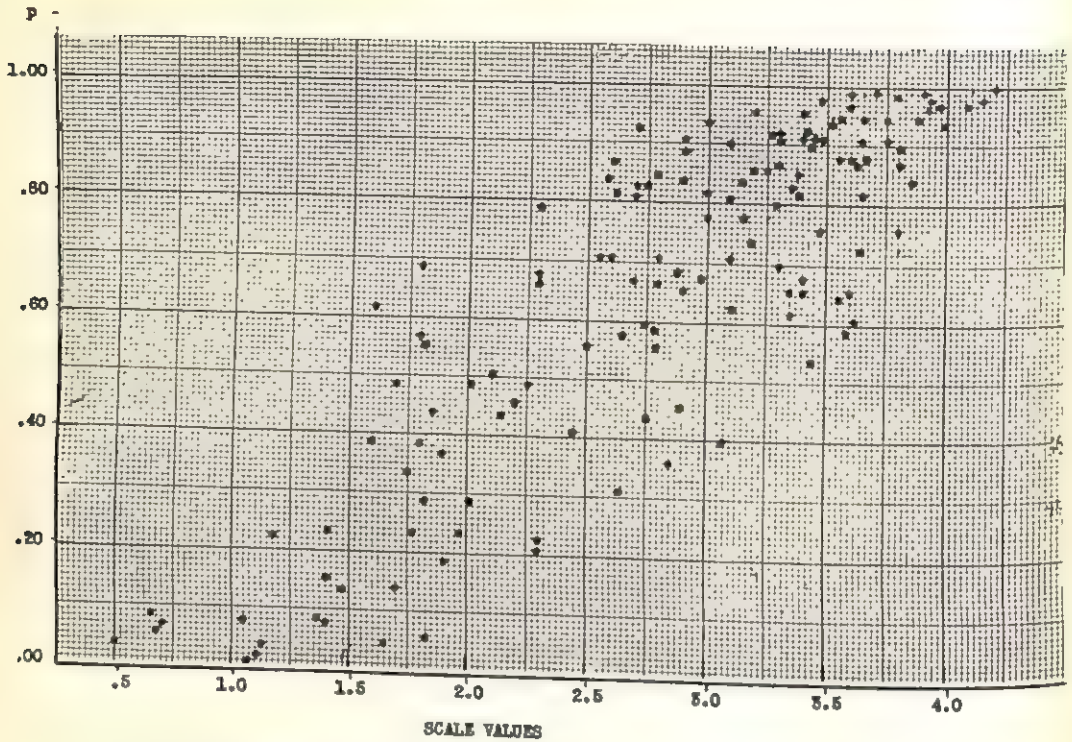


FIG. 2. Probability of endorsement of a trait item plotted against the social desirability scale value of the item. The product-moment correlation coefficient is .871.

by a particular item is characteristic of themselves. The proportions may be taken as the probability of endorsement of a particular trait item for the sample at hand.

The probability of endorsement of each item was plotted against the previously, and independently, determined social desirability scale value of the item. This plot is shown in Figure 2. On the Y-axis we have the probability of endorsement and on the X-axis the social desirability scale value. It is apparent that the probability of endorsement is a linear function of the scaled desirability of the item.<sup>1</sup> The product-moment correlation coefficient is .871.

#### Discussion

The data clearly indicate that the probability of endorsement of an item increases with the judged desirability of the item.

<sup>1</sup> There is a slight indication of departure from linearity at the two extremes of the scale value axis. This is probably because of the limit placed upon the plotted points in terms of the Y-axis. The departure from linearity, however, is not statistically significant.

This does not necessarily mean that the subjects are misrepresenting themselves on the inventory. It may be that traits which are judged as desirable are those which are fairly widespread or common among members of a culture or group. That is, if a pattern of behavior is prevalent among members of a group, it will be judged as desirable; if it is uncommon, it will be judged as undesirable. We might thus expect items indicating desirable traits to be endorsed more frequently than items indicating undesirable traits.

It is also possible that the behavior indicated by an item with a high social desirability scale value is not common, but that the subject taking the inventory is trying, consciously or unconsciously, to give a good impression of himself. He therefore tends to distort his answers in such a way as to make himself out as having more of the socially desirable traits and fewer of the socially undesirable traits than might be the case if his behavior were evaluated in terms of some other independent criterion.

Either one or both of the interpretations presented would account for the relationship between probability of endorsement and scaled desirability of the item. I have no data to support the interpretation that the subjects misrepresented themselves on the inventory, but Ellis (2) in his recent review cites quite a few studies which would indicate that this is the case.

If this is true, then in a personality inventory we should attempt to minimize the tendency for a given response to be determined primarily by the factor of social desirability. A suggested solution is to pair items indicative of different traits in terms of their social desirability scale values. If the subject is

then forced to choose between the two items, his choice obviously cannot be upon the basis of the greater social desirability of one of the items.

*Received June 3, 1952.*

#### References

1. Edwards, A. L. The scaling of stimuli by the method of successive intervals. *J. appl. Psychol.*, 1952, 36, 118-122.
2. Ellis, A. The validity of personality questionnaires. *Psychol. Bull.*, 1946, 43, 385-440.
3. Gordon, L. V. Validities of the forced-choice and questionnaires methods of personality measurement. *J. appl. Psychol.*, 1951, 35, 407-412.
4. Murray, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.

## A Note on "Interest Item Response Arrangement"

John V. Zuckerman

*Human Resources Research Office, The George Washington University*

In a private communication, Cronbach has called my attention to some aspects of my recent article (5) which should be clarified. Some methodological problems were not sufficiently explained in the original article, a basic assumption was left unstated, and in addition some violence was done in citing Cronbach's position with respect to the effect of item response arrangement on measurement of traits or qualities. The points to be considered may be examined topic by topic.

### Reliability

In a comparison of two interest test forms, FE (with 168 two-choice items) and OE (containing 112 L-I-D items), four product-moment reliabilities (corrected for test length by the Spearman-Brown formula) were computed for four empirical keys, using odd-even technique. With one exception, the reliabilities were similar for the two forms. For the key which was discrepant, OE had a higher reliability. Odd-even reliabilities cannot be interpreted as estimates of test-retest reliabilities, particularly because, for L-I-D or similar scales, odd-even figures would be raised in the event that a transient response set were affecting performance throughout the test. Evidence of such response set might be found in the number and direction of weights for responses to the different categories. One may conclude that the split-half correlation is not the appropriate reliability measure for an empirically keyed scale, since a test might have low inter-item consistency but a high test-retest reliability. A low split-half correlation would obscure the high real reliability.

Retabulation of my data in terms of numbers of positions weighted (see Table 1) shows that there are consistent tendencies for educators to like more things than engineers, teachers to dislike more things than administrators, administrators to like more things

than teachers, and teachers to dislike more things than educators in general.

These tendencies were capitalized on by the empirical scoring keys used in the study. To determine whether reliabilities of those L-I-D keys are lower than for forced-choice keys, the study would require the addition of test-retest reliability information which is not presently available.

### Validity

My study was intended to compare the relative discrimination provided by forced-choice and L-I-D item forms. The experimental design involved the assessment of relative discrimination of four scales by rescoring blanks of most of the original subjects. The assumption was made that any shrinkage for OE scales would be the same for FE scales upon a cross-validation. Cronbach points out that when L-I-D or similar three-choice items are assigned weights, there are more possibilities that weights could arise out of chance differences than where forced-choice pairs are used. He states further that the more chance discriminations are counted in the score, the more the validity will shrink on a fresh sample, based on mathematical considerations.

Table 1

Retabulation of Zuckerman's Data (5) in Terms of Numbers of Positions Weighted

Scale Name	No. of Items Weighted	Positions Weighted*					
		L+	L-	I+	I-	D+	D-
ED-ENG	94	66	16	3	39	25	32
ADM	38	11	9	12	5	3	22
TEA	41	1	18	22	1	6	15
AD-TEA	49	25	6	6	15	8	22

\* Positive direction of weights in favor of educators for ED-ENG, for administrators and teachers for the ADM and TEA scales, and for administrators in the AD-TEA scale. See original article for explanation of scale construction (5).

Probable shrinkage, according to Cronbach, depends on factorial complexity of the item matrix, the number of items (or weights), the number of subjects tested, and the criterion reliability. No data are available from my study which bear on the problem of differential shrinkage for different item forms. To settle this point, a follow-up study will have to be made in which cross-validation procedures are used.

In a study intended to achieve the same aims as mine, but with a different methodology and subject matter, Gordon (4) found that forced-choice personality questions provided more discrimination than open-ended rating scale statements. Differences in criteria, measuring instruments, and methods prevent direct comparison of the studies, however.

#### Response Set

While the general tone of Cronbach's earlier articles on response set (1, 2) was unfavorable toward the use of item forms such as L-I-D, he did clearly raise the possibility that it is desirable to capitalize on response-set variance (especially 2, pp. 17, 27, and 28). My statements regarding his position (5, pp. 79 and 84) were in error. A

recent study by Cronbach and another author (3) expresses Cronbach's current appraisal of the problem in terms of a mathematical consideration of profile analysis. Response-set may appear as a mathematical factor entitled *elevation*. The investigator is advised to consider the meaning, if any, of the factor, and determine whether it is to be included in his scoring procedure. There is no basic disagreement between Cronbach and myself on this point.

Received January 21, 1953.

Published out-of-turn by the editor.

#### References

1. Cronbach, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, 6, 475-493.
2. Cronbach, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, 10, 3-31.
3. Cronbach, L. J., and Gleser, Goldine C. *Similarity between persons and related problems of profile analysis*. Technical Report No. 2, ONR Contract N6ori-07135. Urbana, University of Illinois, April, 1952.
4. Gordon, L. V. Validities of the forced-choice and questionnaire methods of personality measurements. *J. appl. Psychol.*, 1951, 35, 407-412.
5. Zuckerman, J. V. Interest item response arrangement as it affects discrimination between professional groups. *J. appl. Psychol.*, 1952, 36, 79-85.

## Effects of the Nature of the Problem on LGD Performance \*

Bernard M. Bass and Cecil R. Wurster

*Louisiana State University*

The basic scheme of group situational tests is to place examinees as a group in a problem or work situation. Examiners observe, record, or rate examinees' behavior as members of the group. The hypothesis underlying the method is that the situational test is a valid sample of behavior for predicting future behavior in a real group situation. It has been verified by a number of studies (e.g. 2, 4, 11).

Because of the wide variety of possible group situations, a large number of variations in group situational tests have been tried. Candidates for positions of leadership have been assessed: (a) in initially leaderless situations (e.g. 4); (b) in situations where each candidate, in turn, has been appointed leader (e.g. 2); (c) in situations where a staff member has served as leader (e.g. 1); and (d) in situations where the leader has been elected by the group (7). Arbous and Maree have reported a median correlation of .67 between assessments based on observations of the same candidates in situations a and b.

A problem for solution may or may not have been presented. Some studies have given participants a choice of problems to discuss (e.g. 11); others (e.g. 9) have allowed the group to originate the problem; while still others have assigned the problem (e.g. 4).

Various kinds of problems have been presented. These have included general interest problems such as "Select the ten outstanding leaders in the world today" (10); more specific problems such as: "Develop a program to train supervisors in this plant" (3); as well as case histories of human relations problems in which the group is asked to decide what the best course of action will be (8).

\* This study was aided by a grant from the Louisiana State University Graduate Council on Research. The writers wish to express their appreciation to Mr. Ernest McNeil, Mr. Jamie Dennis and the many others whose help made this study possible.

The purpose of the present study was two-fold. The first aim was to see the extent to which a person's successful leadership activity in an initially leaderless discussion changed when there was a systematic change in the nature of the problem and the persons with whom he was grouped. The second purpose was to see whether assessments based on some types of discussion situations were more related than others to various measures of company rank, education, intelligence, supervisory aptitude, age and appraisals of supervisory behavior on-the-job.

### Subjects and Method

The subjects were a class of 23 students in an introductory psychology course and 131 oil refinery supervisors. The 23 students were divided purposefully into three groups and each observed in a half-hour LGD with one of three types of conditions: (a) unstructured—participants originated problem for discussion; (b) general leader specifications—e.g. participants developed a set of factors for choosing the world's greatest leaders; (c) case history—e.g. participants decided whether a returning veteran should tell his wife about an illegitimate child he fathered overseas.

Then, three new groups were formed so that as few members of the same first three groups were together for a second time and so that all participants could be assessed under a condition different from the situation in which they were first tested. Finally, a third recombination was carried out so that all 23 participants were observed under each of the three conditions. Conditions b and c were altered slightly on each successive administration to avoid having participants specifically prepared. Thus, different kinds of specifications were demanded and different case histories were used in the successive administrations. LGD scores were based on

Table 1  
Correlations Among LGD Scores Earned by Participants Subjected to  
Three Different Types of Discussion\*

Type of LGD	Type of LGD			Average
	Unstructured	Leader Specifications	Case History	
Unstructured	—	.58	.66	.62
Leader Specification		—	.51	.54
Case History			—	.58

\*  $N = 23$ .

one observer's ratings of the extent to which each participant exhibited successful leadership activity in a given discussion.<sup>1</sup>

The 131 supervisors were assessed under one of four conditions: (a) unstructured; (b) general leader specifications; (c) in-plant leader specifications; (d) case history. Situation c concerned the specifications for selecting shift foreman, supervisors and so forth. The case history concerned such problems as what Mike should do when his superior bawls him out in front of his subordinates or what Harry's superior should do when he finds various faults with Harry's method of leading a work gang. Five unstructured and four of each of the other types of situations comprised the 17 group discussions. LGD scores were corrected for group size and variations among observer's standards.<sup>2</sup>

### Results

Table 1 displays the intercorrelations among LGD scores obtained by the 23 participants on the basis of each of the three types of discussions.

Since an LGD test-retest reliability of .75 was reported for repeated discussions a week apart with changed participants but with no change of problem (8); and since it has been found even lower ( $r = .53$ ) when a year intervenes between repeated measurements and the outside status of some participants is changed more than others (5), it was inferred from the results reported in Table 1 that some, but not very much, variation in

LGD behavior could be attributed to variations in the nature of the stimulating situation.

The variation from .51 to .66 in intercorrelation and from .54 to .62 in average intercorrelation are most probably due to chance. Since the usual validities of these various types of LGD's range from .30 to .50, these intercorrelations suggest that to include more than one in a battery would not raise the validity very much of any two over the validity of any one, although the reliability of the composite might be raised substantially.

Table 2 indicates the correlations between two independent clusters of highly interrelated variables and LGD scores earned by the refinery supervisors in one of four types of discussions. (The clusters were isolated by inspection of an intercorrelation matrix. Cluster I consisted of supervisor's rank in the company, education, intelligence, supervisory aptitude and youth. Cluster II consisted of superiors on-the-job appraisals of the supervisors by means of graphic and forced-choice rating scales (6).) Supervisors were classified into a lower and upper echelon of management. Correlations between rank and LGD scores were biserial; the remaining were Pearson product-moment.

Chi square tests of the significance of the variations in correlations from one discussion type to the next<sup>3</sup> suggested that only one set of correlations—those between company rank and LGD scores—varied significantly at the 1 per cent level of confidence. It was in-

<sup>1</sup> For a more detailed description of the scoring procedure, the reader is referred to (6).  
<sup>2</sup> See footnote 1.

<sup>3</sup> This test is described by Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950. Pp. 133-135.

Table 2

The Correlation Between LGD Scores of Oil Refinery Supervisors Subjected to One of Four Types of Discussion Situations and Their Rank, Education, Intelligence, Supervisory Aptitude, Age and Superior's Appraisals

	Sub-Samples According to LGD Type				Total
	Unstructured (a)	Out-Plant Leader Specifications (b)	In-Plant Leader Specifications (c)	Case History (d)	
No. of Groups	5	4	4	4	17
No. of Subjects	35*	33*	31*	32*	131
Cluster I					
Rank†	.87	.81	.91	.90	.88
Education	.63	.47	.60	.57	.57
Intelligence	.50	.61	.41	.34	.45
Supervisory Aptitude	.34	.28	.07	.54	.30
Youth	.30	.01	.31	.24	.19
Cluster II					
Graphic Appraisal	.02	-.38	-.11	-.04	-.01
Forced Choice Appraisal	.02	-.22	-.12	.28	-.12

\* Because of missing information on intelligence, supervisory aptitude scores and superior's appraisals many of the sub-sample correlations of LGD scores with these variables are based on as few as 21 cases.

† Sub-sample variation in correlation with LGD significant at the 1 per cent level of confidence. This set computed by means of biserial  $r$ . The others are Pearson correlations.

ferred from the correlation of .99 between rank and case history LGD scores that upper echelon supervisors were the sole leaders in such discussions; their tendency to exert leadership in the supposedly leaderless situation declined somewhat when discussions involved situations outside the company as in the out-plant leader specifications and the unstructured discussions. (In the latter, the problem originated by the participants for discussion quite often concerned improving the town sewerage system, increasing civic pride, and so forth.) The one hypothesis worthy of further investigation drawn from these results, therefore, was that a supervisor of high rank is most likely to play the role of leader among persons of lower appointed rank when the group problem specifically concerns situations for which he has the high rank.

In evaluating these results, the reader should note that as reported previously (6), there were very large restrictions in the range of most of the variables—especially, supervisory aptitude scores and superior's appraisals—because a large percentage of the examinees were selected for their present

posts because of their high supervisory aptitude test battery scores.

The correlation of .54 between "case history" LGD scores and supervisory aptitude suggested a valid consistency between assessments based on the case history LGD and the supervisory aptitude battery—a battery which gave substantial weight to paper-and-pencil tests of supervisory judgment. When the masking influence of rank, the lower reliability and validity of the graphic in comparison to the forced choice appraisal and the great restriction in range of the appraisals were all taken into account, it was inferred from the correlation of .28 between case history LGD scores and forced choice appraisals that the case history LGD is the most likely type of those investigated to provide a valid predictor of adequacy on-the-job, where the examinees are of different known organizational rank, and previously have been selected by means of valid paper-and-pencil test batteries.

### Summary

The purposes of the present study were to see the effects on their behavior of changing

the nature of the problem confronting LGD participants.

LGD scores of 23 college students correlated between .51 and .66 with repeated administrations where the composition of the group and the problem for discussion were systematically altered. These correlations were not much lower than the test-retest reliability ( $r = .75$ ) of one type of LGD.

The extent to which various personal factors were associated with LGD performance of 131 oil refinery supervisors depended to some extent on the nature of the problem under discussion. Major findings were:

1. A high-ranking supervisor is more likely to exert leadership in small discussion groups with supervisors of lower rank when the discussion specifically concerns situations for which he has the high rank.

2. The amount of successful leader activity in discussions of case histories of human relations problems appears related to paper-and-pencil predictors of supervisory success ( $r = .54$ ) and to a lesser extent with forced choice on-the-job appraisals of supervisory success ( $r = .28$ ).

Received June 2, 1952.

## References

1. Ansbacher, H. L. Lasting and passing aspects of German military psychology. *Sociometry*, 1949, 12, 301-312.
2. Arbous, A. G., and Maree, J. Contribution of two group discussion techniques to a validated test battery. *Occup. Psychol.*, 1951, 25, 1-17.
3. Bass, B. M. Situational tests. II: Variables of the leaderless group discussion. *Educ. psychol. Measmt.*, 1951, 11, 196-207.
4. Bass, B. M., and Coates, C. H. Forecasting officer potential using the leaderless group discussion. *J. abn. soc. Psychol.*, 1952, 47, 321-325.
5. Bass, B. M., and Coates, C. H. Studies of leadership in ROTC. In preparation.
6. Bass, B. M., and Wurster, C. R. Effects of company rank on LGD performance of oil refinery supervisors. *J. appl. Psychol.*, 1953, 37, 100-104.
7. Fields, H. An analysis of the group oral interview. *Personnel*, 1951, 27, 480-486.
8. French, R. L., and Bell, B. Consistency of individual leadership position in small groups of varying membership. *J. abn. soc. Psychol.*, 1950, 45, 764-767.
9. Garforth, G. I. De La P. War officer selection boards. *Occup. Psychol.*, 1945, 19, 97-108.
10. Taft, R. *Some correlates of the ability to make accurate social judgments*. Ph.D. Dissertation, University of California: Berkeley, 1950.
11. Wurster, C., and Bass, B. M. Situational tests: IV. Validity of leaderless group discussions among strangers. *Educ. psychol. Measmt.* In Press.

## Effects of Company Rank on LGD Performance of Oil Refinery Supervisors \*

Bernard M. Bass and Cecil R. Wurster

*Louisiana State University*

Mandell (4), among others, has hypothesized that candidates for employment or promotion who are assessed in a leaderless group discussion or group oral performance test should be unacquainted with each other; otherwise "they may defer to a candidate who has high prestige in the group, or who has a higher-level position." The primary purpose of this study was to investigate the extent to which a person's performance in the LGD was influenced by his administrative rank outside the immediate stimulating situation.

A number of sub-hypotheses were tested and a number of relationships were uncovered concerning the interactions between company rank, degree of successful leader activity in the LGD, rated performance by superiors as a supervisor, age, education, intelligence and knowledge and attitudes predictive of success in supervisory work.

It was believed that the results would be of interest to those engaged in using the LGD to screen applicants for employment or promotion. They would also provide further information to a growing body of knowledge concerning leader-follower relations in small groups.

### Subjects

A total of 131 supervisors at a large oil refinery participated in leaderless group discussions. Of these, 61 were first level maintenance department supervisors; 22 were first level process and production supervisors; 18 were second level and 7 were third and fourth level supervisors in production, maintenance, or staff positions. In addition, 20 were engineers, accountants or other technicians who had no supervisory positions while three had

highly responsible technical positions which called for little direct supervision. The subjects ranged in age from under 30 to over 60 and from sixth grade to Ph.D. in education. The average subject was 43 years old and a high school graduate.

One restriction which most probably served to severely attenuate the various relationships studied was caused by the high percentage of subjects who had been selected for their jobs by a previously-validated battery of psychological tests. A further factor which probably served to restrict the range of observable differences was the large amount of supervisory training these subjects had received from various formal and informal programs.

### Method

Approximately 20 supervisors at a time met for a week-long supervisory training program. On the fourth day, they were subdivided into two or three groups, 6, 7, 8, 9, or 10 to a group, and administered one of four types of leaderless discussions.<sup>1</sup> A total of 17 discussions was run, each observed by one of four trained raters. Directions and scoring<sup>2</sup> were similar to previous studies at Louisiana State University (e.g. 2).

Two types of criteria of on-the-job success as supervisors were available: forced-choice and

<sup>1</sup> A separate report will deal with variations in performance on the LGD as a function of the nature of the discussion problem.

<sup>2</sup> The single observer rated on a 5-point scale the extent to which each participant exhibited the following 7 behaviors: (1) showed initiative; (2) spoke effectively; (3) clearly defined problem; (4) offered good solutions; (5) influenced others; (6) motivated others; (7) led the discussion. A participant's LGD score was the sum of points he received, corrected for group size, and observer variations in points assigned. Scores were adjusted according to the mean score earned by participants of groups of a given size. The distribution of scores assigned by each observer was transformed into a sten distribution (3) in order to make fairly comparable all adjusted scores assigned by the different observers.

\* This study was aided by a grant from the Louisiana State University Graduate Council on Research. The writers wish to express their appreciation to Mr. Ernest McNeil, Mr. Jamie Dennis, and the many others whose help made this study possible.

graphic appraisals by the subjects' superiors. Forced-choice supervisory performance report ratings for 1950 by at least two of their superiors were obtained from the records of 123 of the subjects. These ratings, developed by Richardson, Bellows and Henry, Inc., had odd-even and equivalent-form reliabilities above .90 for the groups on which they were standardized. Inter-rater reliability was .69. Validity of the ratings as measured by their tendency to differentiate previously identified, above average, average and below average supervisors ranged from .62 to .84 for the various forms and departments of the refinery. However, for the present restricted sample, inter-rater agreement was only .43. Since most subjects' average appraisals were based on independent ratings by as many as six superiors, the actual reliability of this measure was appreciably higher.

Corresponding graphic ratings were likewise available for these 123 subjects. For this sample, inter-rater correlation was only .29. Otis Intelligence Test scores for 87 subjects and "supervisory aptitude" test scores for 92 subjects were also available. The supervisory battery test scores were an optimally weighted sum of scores of performance on a forced-choice test of supervisory judgment, an empirically scored forced-choice personality inventory, and scores on certain keys of the Kuder Preference Record. For the original standardizing group, the optimally weighted battery of scores correlated .62 with superiors' ratings of the subjects.

### Results

Table 1 displays the matrix of intercorrelations among LGD scores, company rank, education, intelligence, supervisory aptitude, youth,<sup>3</sup> forced-choice appraisals and graphic superior's appraisals. The first six have been grouped into one cluster of highly intercorrelated variables while the last two form a second cluster. The average correlation between each variable and all the others of cluster I and cluster II are also shown. All correlations reported are Pearson product-moment except those between rank and the other variables which are biserial.<sup>4</sup>

Criterion ratings, supervisory battery scores and intelligence test scores were available in

standard score form with means of 20 and standard deviations of 5 for the original standardizing population of supervisors and candidates for supervisory positions. It should be noted that the sample used in this study was decidedly restricted in range on these significant variables. The sample mean was half a standard deviation higher in mean criterion ratings and supervisory aptitude scores than the original population from which many of its members were drawn. Restrictions in range were from 12 to 58 per cent which severely attenuated the relationships reported.

The first cluster of six variables had a mean intercorrelation of .48 while this cluster correlated .00 with the second cluster of two variables. Thus, it appeared that performance on the LGD was highly related to company rank ( $r_b = .88$ ) and to a lesser extent with the other variables closely associated with rank: education ( $r = .57$ ), intelligence ( $r = .45$ ), supervisory aptitude ( $r = .30$ ) and youth ( $r = .19$ ). LGD performance was unrelated to superiors' appraisals. Further analyses indicated that the mean LGD scores (in tens) for subjects from the first, second and combined third and fourth echelons of supervision were 3.6, 6.7, and 6.7 respectively which according to an analysis of variance were significantly variant at the 1% level. No such significant differentiation was found when all first-line maintenance supervisors whose mean LGD score was 3.4 were compared with all first-line process and production supervisors whose mean was 4.0. Staff and technical men, not included in the above samples, had a mean LGD score of 5.4. This intermediate value reflected probably their subordinate position compared to upper echelon supervisors but their superior education and intelligence to first-line supervisors.

Company rank appeared significantly related to forced-choice criterion ratings earned ( $r_b = .34$ ) but not to graphic ratings. Rank correlated significantly with supervisory battery test scores ( $r_b = .42$ ). In this highly technical industry, it was not surprising to observe the almost complete interdependence of supervisory rank and education ( $r_b = .98$ ).

<sup>3</sup> Age reversed in sign to make positive most of the correlations of age with the variables of cluster I.

<sup>4</sup> The small proportion of supervisors in the second, third and fourth echelons of management included in this study, led the investigators when correlating rank with the other variables to combine them into one upper management group of 25 cases to compare with one first level supervisory group of 83 cases.

Table 1

Intercorrelations Among Company Rank, Education, Intelligence, LGD Score, Supervisory Aptitude, Youth, Superiors' Appraisals, and Two Clusters of Highly Intercorrelated Variables

	Cluster I					Cluster II		Average Correlation with		
	Com- pany Rank	Edu- cation	Intelli- gence	LGD Score	Super- visory Aptitude	Youth	FC Appraisal	Graphic Appraisal	Cluster I	Cluster II
Company Rank		.98*	<i>Inc.</i>	.88*	.42*	.44*	.34*	.07	.68	.20
Education			.57*	.57*	.31*	.32*	.03	-.22†	.55	-.10
Intelligence				.45*	.56*	.43*	.09	-.18†	.50	-.04
LGD Score					.30*	.19	-.01	-.12	.48	-.06
Supervisory Aptitude						.29*	-.01	-.08	.38	-.04
Youth							.20†	.06	.33	.07
Forced-Choice Appraisal								.68*	.11	.68
Graphic Appraisal									-.10	.68
Mean		12.0	20.4	4.5	22.5	42.9	22.6	22.8		
Standard Deviation		3.3	4.4	2.0	3.4	8.4	3.2	2.1		

\*  $P < .01$ .

†  $P < .05$ .

*Inc*—Not enough data on intelligence available in upper echelons to compute correlation.

Italicized coefficients are based on biserial  $r$ ; the remainder are Pearson product-moment correlations.

Not enough cases were available to obtain the correlations between rank and intelligence although it is expected that it was at least .50 since education and intelligence in this sample correlated .57.

To see if company rank was masking any relationships between the other variables, two attempts were made to study the relations among the other variables when rank was partialled out. Table 2 shows the appropriate partial correlations among LGD scores, supervisory aptitude, youth and forced-choice and graphic appraisals. Since education and rank were about perfectly correlated,

it was impossible to partial out one without eliminating the variance of the other.

When rank is partialled out, a large percentage of variance in the amount of successful leader activity is accounted for by youth ( $r_{01.2} = -.70$ ). At the same time, men rated as inadequate by their superiors on the forced-choice performance reports ( $r_{01.2} = -.69$ ) and the graphic ratings ( $r_{01.2} = -.38$ ) tend to attain high LGD scores. Correlations among the other variables remain unaffected by partialing out rank or else are reduced to negligible importance.

There was some doubt about the meaning

Table 2

Partial Correlations Between LGD Score, Supervisory Aptitude, Youth, and Superiors' Appraisals, with Company Rank Held Constant Statistically

	LGD Score	Supervisory Aptitude	Youth	FC Appraisal	Graphic Appraisal
LGD Score					
Supervisory Aptitude		-.16	-.70	-.69	-.38
Youth			.13	-.17	-.12
Forced-Choice Appraisal				.06	-.10
Graphic Appraisal					.70

of these partial correlations. First, in order to use partial  $r$ , it was necessary to assume that the biserial  $r$  correlations between rank and the other variables were estimates of the product-moment correlations between rank and the other variable. Second, partial  $r$  estimates the amount of the relationship between a pair of measures ruling out the effects of a third *when the remaining variances of the pair of measures are equal*. But, if company rank is held constant, it will impose different restrictions in range on each member of a pair of variables so that partial  $r$  provides a description which usually never exists in reality. Thus, if the variances of forced-choice appraisal and LGD scores could be equalized after company rank was held constant, then a correlation of  $-.69$  would exist between them. But it is seldom that this equalization of variance occurs in nature. Therefore, a second approach—purposive sampling—was used to rule out the effects of company rank on the correlations between LGD scores and the other variables. Table 3 shows these correlations for first level supervisors only and for upper level supervisors only. From the results in Table 3, it was inferred that when rank is held constant, experimentally, the correlations between LGD performance and the other variables tend to reduce to insignificance. This was not unexpected since LGD score and rank correlated  $.88$ .

As shown in Table 3, first-line supervisors with high LGD scores tended more than

upper echelon supervisors to be considered inadequate on-the-job, although the differences were not significant. The less extreme results obtained through this sampling procedure as compared with the partial  $r$  approach was attributed to the fact that no attempt was made to equalize the variance of each two variables correlated in the upper and lower supervisory ranks while partial  $r$  forced such equalization.

### Conclusions

The results of this study are a strong confirmation of a number of common-sense observations as well as research findings about the influence of a person's rank, prestige or status in an organization and his tendency to play the role of leader in small groups of members from that organization even where there is no appointed leader for the immediate situation.

The biserial correlation of  $.88$  between a participant's company rank and his leader behavior in a supposedly initially leaderless discussion is consistent with a number of other studies. For example, Bass and Coates (1) found that there was a significantly greater increase in LGD scores on a retest a year after the original test by ROTC cadets who had been promoted to positions of cadet first lieutenant or higher during the period which intervened between test and retest than the remainder (who had become cadet second lieutenants). Similarly, Michigan Conference Research studies suggest that

Table 3  
Correlations Between LGD Score and Supervisory Aptitude, Youth, and Superiors' Appraisals,  
with Company Rank Held Constant by Purposive Sampling

Variable	Company Rank			
	First Level		Second, Third or Fourth Level	
	N	$r$ with LGD Score	N	$r$ with LGD Score
Supervisory Aptitude	69	.29†	6	.21
Youth	83	.09	25	.02
Forced-Choice Appraisal	81	-.12	23	-.04
Graphic Appraisal	77	-.20	22	-.12

†  $P < .05$ .

when three-man appraisal boards meet, the conclusions reached are those in agreement with the member of highest status. Also executives appear to call so-called planning conferences of their subordinates mainly to obtain subordinates' agreement on what the executive has already decided to do (5).

In previous studies of unacquainted candidates or candidates of similar initial rank there have uniformly been reported, by at least 11 separate investigations, correlations ranging from .30 to .70 between LGD scores and various criteria of supervisory success or leadership potential. In the present study, these correlations were close to zero suggesting strongly that Mandell's suspicions are confirmed concerning the general lack of validity of the LGD among acquaintances, especially where they differ greatly in initial prestige or rank.

The theoretically significant negative correlations between LGD scores and criteria of supervisory success when company rank is partialled out statistically, pose more questions than they answer. These include:

1. Is one of the requirements necessary to be a successful first-line supervisor, the ability to play a subordinate role when in a social situation with those of higher company rank than he, even though they are not his immediate superiors and the situation is outside plant jurisdiction? Or, on the other hand, are organizations discouraging communication upward from lower echelon management as well as hindering executive development by appraising as inadequate those first-line supervisors who give suggestions, opinions and information, who take initiative and show originality in their interactions with their superiors?

2. Is it the younger, less secure supervisor who is most conscious of rank and least apt to ignore it, even in unstructured social situations? If so, what effect does this have on the introduction of new ideas into an organization?

3. Since active trainees, trainees who receive and take advantage of the opportunity to make decisions, usually learn more than passive ones, to what extent are first-line supervisors handicapped when placed in conference training with upper-echelon personnel?

### Summary

LGD scores of 131 oil refinery supervisors were correlated with their rank in the refinery, their education, intelligence, "supervisory aptitude" test scores, and supervisors' appraisals of their on-the-job performance. LGD scores correlated .88 with rank, .57 with education; .45 with intelligence, .30 with supervisory aptitude and -.19 with age. Most of these correlations could be attributed to the influence of rank on all these variables. When rank was partialled out statistically, LGD scores were highly positively related to age and highly negatively related to superiors' appraisals. It was concluded that in general the LGD is not valid where participants are of known different rank. The complexity of the outcomes of this study raise some interesting questions about the validity of superiors' appraisals as ultimate criteria of supervisory performance and the influence of formal rank on the behavior of conference participants of differing rank.

Received May 19, 1952.

### References

1. Bass, B. M., and Coates, C. H. . Studies of leadership in ROTC. In preparation.
2. Bass, B. M., McGhee, C. R., et al. Personality variables related to leaderless group discussion behavior. *J. abnorm. soc. Psychol.*, 1953, 48, 120-128.
3. Canfield, A. A. The "Sten" Scale—A modified C-Scale. *Educ. psychol. Measmt.*, 1951, 11, 295-297.
4. Mandell, M. The group oral performance test. Washington: U. S. Civil Service Comm., April, 1952.
5. Conference Research, University of Michigan. *Process of the Administrative Conference*. Contract N6 onr-232, T.O.VII, March, 1950.

## Flesch Readability Analysis of the Major Pre-election Speeches of Eisenhower and Stevenson

Arthur I. Siegel

*Institute for Research in Human Relations, Philadelphia*

and

Estelle Siegel

*Drexel Hill, Pennsylvania*

One non-political point of interest during the recent election campaign was the level at which each of the rival presidential candidates was speaking. For instance, some people maintained that Stevenson was doing himself an injustice because he was speaking over the heads of his audience, e.g., he was being too scholar-like, pedantic, academic, formal, learned, etc. On the other hand, some chided the same candidate for being a joker or a punster. In order to gain some insight into the legitimacy of these arguments, a Flesch readability analysis of the texts of six of the major talks of Stevenson was performed. For comparative (control?) purposes, the texts of the major talks given by Eisenhower on the same dates were also analyzed.

### Method

The texts of the major talks of each of the rival candidates appeared in the *Philadelphia Inquirer* on the morning following the talks. In some cases the newspaper saw fit to delete certain parts of the speeches of each candidate. In these cases, only the "selected text" was available for analysis. On October 28, Eisenhower's major appearance was a television show, in which Eisenhower answered questions posed by various Republican committee women. In this case, the text of the president-elect's replies to these questions was analyzed. Originally, it was our intent to analyze the texts of the six talks of each candidate given just prior to the election. However, the Sunday, November 2, issue of the *Inquirer* did not contain the previous day's talks of either candidate. Neither candidate spoke on Sunday, November 2. Thus, the texts of the major talks of Eisenhower and

Stevenson on October 27, 28, 29, 30, 31, and November 3 were included in the present study.

Flesch<sup>1</sup> recommends, as a sampling procedure, that every third paragraph be taken and that the first 100 words of each sampled paragraph be analyzed. However, since many of the paragraphs of each candidate ran under 100 words, the entire texts were analyzed.

### Results

The results of the analysis are presented in Table 1.

The "reading ease" of the texts of three of the major talks given by Stevenson during the final eight days of the campaign were classified as "Standard" by the Flesch analysis, and three were classified as "Fairly difficult." For the same period, the reading ease of the texts of four of Eisenhower's talks were classified as "Standard," while one was classified as "Difficult," and one was classified as "Fairly difficult." The mean reading ease score of Eisenhower's speeches was "Fairly difficult" and of Stevenson's was "Standard." But the actual difference of only 1.5 points is negligible. A "Standard" style of reading ease is found, according to Flesch, in *Digests*; a "Fairly difficult" style is characteristic of academic publications.

From the standpoint of "human interest," the Flesch analysis indicated the style of the texts of four of Eisenhower's speeches to be "interesting" and two to be "highly interesting." The styles of five of Stevenson's speeches were "interesting" and one was "highly interesting." An "interesting" style,

<sup>1</sup> Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.

Table 1

Flesch Reading Ease and Human Interest Scores and Descriptions for Texts of Six Pre-election Talks by Eisenhower and Stevenson

Reading Ease				
Date	Eisenhower		Stevenson	
	Score	Description	Score	Description
Oct. 27	53	Fairly difficult	61	Standard
Oct. 28	60	Standard	57	Fairly difficult
Oct. 29	66	Standard	59	Fairly difficult
Oct. 30	63	Standard	66	Standard
Oct. 31	46	Difficult	55	Fairly difficult
Nov. 3	65	Standard	64	Standard
Mean	58.8	Fairly difficult	60.3	Standard
S.D.	4.1		3.8	

Human Interest				
Date	Eisenhower		Stevenson	
	Score	Description	Score	Description
Oct. 27	28	Interesting	41	Interesting
Oct. 28	39	Interesting	28	Interesting
Oct. 29	51	Highly interesting	42	Highly interesting
Oct. 30	43	Highly interesting	30	Interesting
Oct. 31	33	Interesting	39	Interesting
Nov. 3	27	Interesting	34	Interesting
Mean	36.8	Interesting	34.0	Interesting
S.D.	8.5		5.6	

according to Flesch, is found in the *Digests*, while a "highly interesting" style is found in the *New Yorker*.

Thus, using the Flesch analysis as a yardstick, for the period investigated, we have little evidence to indicate that Stevenson approached the academic level, nor were his

speeches more difficult to understand than Eisenhower's. On the other hand, by a Flesch analysis, there was a slight tendency for Eisenhower's speeches to be more "interesting."

Received January 29, 1953.  
Early publication.

# Factorial Analysis of the Original and the Simplified Flesch Reading Ease Formulas<sup>1</sup>

Marvin D. Dunnette  
*Industrial Relations Center,  
University of Minnesota*

and

Paul W. Maloney<sup>2</sup>  
*The Addison Lewis Co.,  
Minneapolis*

Farr, Jenkins, Paterson, and England (5) have presented evidence showing that the simplified reading ease formula by Farr, Jenkins, and Paterson (4) yields scores quite in agreement with those obtained with the original Flesch formula (6). The Flesch formula was simplified in order to provide a method which "would obviously be much faster and would require no knowledge of syllabification on the part of the analyst" (4, p. 333). The presentation of the simplified formula, however, was not greeted with universal acceptance. Klare (8) and Flesch (7) raised two principal objections. First, they doubted the claimed time economy of the new method. This argument was met with a study by the Minnesota group (5) in which a number of graduate students determined reading ease scores by both methods. The new method was found to be much faster.

But their second objection to the new method is more formidable, and has as yet been unanswered. The argument states that counting one syllable words is less accurate than counting syllables. The logical basis for the position is sound—syllable counting involves attentive study of each word, where-

curacy of the two counting methods. Other aspects of reading ease calculation were also investigated.

## Method

Untrained<sup>3</sup> subjects for the experiment included 72 male and 72 female freshman students. All were enrolled in freshman English in the School of Agriculture and Home Economics at the University of Minnesota.<sup>4</sup> These students had never been trained in and probably had never heard of the techniques employed in making readability analyses.

Since the number of syllables is inversely related to the number of one syllable words, it became necessary to control the difficulty of the test material.<sup>5</sup> Further, it was felt that ability to perform readability counts may be related to a person's reading ability. Because of this, the subjects were grouped into four relatively homogeneous groups on the basis of their paragraph comprehension scores on the Nelson-Denny Reading Test. The time taken to perform the counts was also measured. Subjects within each group were then randomly assigned to conditions imposed by the following factorial design:

Difficulty Level	Count	Reading Ability Level Group			
		Lower 25%	25%-50%	50%-75%	Upper 25%
Easy	Syllables				
Material	O S Words				
Medium	Syllables				
Material	O S Words				
Difficult	Syllables				
Material	O S Words				

as a person picking out one syllable words might just scan the passage.

This study was designed to bear on the question. A comparison was made of the ac-

<sup>1</sup> Factorial design experiments and the basic computations involved are discussed by Edwards (3) and Nelson (9).

<sup>2</sup> Formerly research assistant in the Industrial Relations Center.

<sup>3</sup> Farr, Jenkins and Paterson (4) have stated that their simplification should remove the aura of complexity from the Flesch formula and make it more useful to practical men in their daily work. Because of this, we felt it would be most desirable to cause of this, we felt it would be most desirable to use naive or untrained subjects who were not oriented in favor of either method of counting.

<sup>4</sup> The authors wish to express thanks to Professor Ralph G. Nichols and Professor James I. Brown who offered the cooperation of their department.

<sup>5</sup> The test material was carefully chosen so as to

A test form was developed for each of these six conditions. The first page of each was a simple explanation of the subject's task. Explanation was facilitated by the use of two short examples which were used for all six conditions. The second page included a 50-word practice passage of the same difficulty as the test passage. The third page was devoted to two test passages of 100 words each. The subjects were instructed to perform the proper count separately for each passage and record their answers in the spaces provided.

Oral instructions were framed to emphasize accuracy over speed. Subjects were asked to check their completed work, and then to record the letter appearing on the blackboard. A new letter was placed on the blackboard every 10 seconds, thus providing time scores without emphasizing the speed factor.

Four factorial designs as shown above were used in the experiment to include information separately for male and female subjects, and for error and time scores. Since 72 subjects of each sex were available, plans provided for three replications in each cell. Several students were absent on the day of the administration. These were immediately sent a test form by mail. In all, 24 students were absent; 20 were located; 16 returned completed test forms. These mailed returns, of course, did not include a time score. It was, therefore, necessary to reduce the number of replications in the two time-score designs to two per cell. Thus, error data for five boys and three girls were missing.

Since statistical analyses of the experimental data required an equal number of replications per cell, the missing scores were estimated. Estimates were based on the best information available; i.e., the scores of the two subjects in the same cell as the missing value. The mean of the two scores was used. The degrees of freedom for error were reduced in accordance with the suggestion of Cochran and Cox (2, p. 73).

Homogeneity of variance was tested by means of Bartlett's Test (3). In no case was Chi-square sufficiently large to reject the hy-

cover the entire range of difficulty. "Easy" registered above 90; "medium" in the 40's; "difficult" below 10. Passages selected registered nearly the same ( $\pm 5$ ) RE scores on both formulas.

pothesis of equal variances. Plotting the data indicated the distributions to be non-skewed and platykurtic. Since skewness is the most serious deviation from normality (for purposes of variance analysis), and since Cochran (1) has shown that non-normality does not seriously alter conclusions derived from variance analysis, no test was made of the assumption of normally distributed parent population.

## Results

Table 1 shows the results of the analysis of variance. It is seen that some of the sources of variation are significantly different from the variation due to error. These results may be interpreted more easily, however, by referring to Table 2 which gives the means and standard deviations for accuracy and time required.

Both boys and girls performed the one syllable word count more accurately than the syllable count. For the boys, the difference was significant at the 5 per cent level. For the girls, however, the difference was not statistically significant. Both boys and girls also did the one syllable word count in 25 per cent less time than that required for the syllable count. The differences were statistically significant.

These findings suggest that the new formula is superior with respect to both accuracy and time required to perform the counts. Since the subjects were unfamiliar with the counting methods required by readability formulas, we can state with assurance that the F, J, and P simplified formula is inherently easier to perform. This finding, combined with previous evidence (5), shows that persons will perform the new count more rapidly and with greater accuracy regardless of the degree of their skill.

Data in Table 1 show a significant interaction for boys between type of count and difficulty of material. Table 3 shows this interaction effect more clearly. The one syllable word count was less accurately performed for easy material but was more accurately performed for difficult material. Expressing the error as a percentage may cover up the practical significance of these differences. This is

Table 1  
Results of the Application of Variance Analysis

Source of Variation	Per Cent Error						Time Taken					
	Boys			Girls			Boys			Girls		
	d/f	M.S.	F	d/f	M.S.	F	d/f	M.S.	F	d/f	M.S.	F
Between Difficulty Levels	2	136.70	2.21	2	103.67	1.15	2	31956.25	3.41*	2	14856.78	.95
Between Reading Ability Levels	3	84.44	1.37	3	171.48	1.90	3	8005.55	.85	3	20488.00	1.31
Between Types of Count	1	362.30	5.87*	1	22.22	.25	1	124033.34	13.24**	1	97651.00	6.23*
Difficulty × Ability	6	59.81	.97	6	125.43	1.39	6	14353.47	1.53	6	8615.07	.55
Difficulty × Count	2	372.66	6.03**	2	181.88	2.01	2	31527.08	3.36*	2	14446.70	.92
Ability × Count	3	31.33	.51	3	75.69	.84	3	6161.33	.66	3	2633.67	.17
Difficulty × Ability × Count	6	36.97	.60	6	40.92	.45	6	4079.75	.44	6	4835.27	.31
Error	43	61.77	—	45	90.44	—	24	9368.75	—	24	15660.92	—

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

Table 2  
Mean Per Cent Error and Mean Time for Different Methods of Counting and for Materials of Different Difficulty

	Boys			Girls			Boys			Girls		
	N	Mean Per Cent Error	S.D.	N	Mean Per Cent Error	S.D.	N	Mean Time (Sec.)	S.D.	N	Mean Time (Sec.)	S.D.
Syllables	32	-4.47	6.84	35	-2.65	7.55	24	380	113	24	366	120
O S W	35	+0.08	9.16	34	-1.43	11.18	24	279	96	24	276	101
Easy	22	-3.49	6.17	23	-3.55	5.67	16	285	99	16	300	103
Medium	22	+0.68	8.05	24	-2.92	11.50	16	331	100	16	356	124
Difficult	23	-3.42	9.92	22	+4.49	9.88	16	374	125	16	308	122

Table 3

Mean Per Cent Error Made by Boys as Related to Difficulty Level and Type of Count

Difficulty of Material	Mean Per Cent Error	
	Syllables	One Syllable Words
Easy	-0.55	-1.67
Medium	-0.58	0.90
Difficult	-3.38	0.76

<sup>1</sup> Data are not included for girls since for them the effect was not statistically significant.

because there are more total syllables than one syllable words in any given reading passage. Therefore, because RE scores depend on the absolute number of units counted, a given variation in reading ease score reflects a larger *per cent* error in the one syllable word count than in the syllable count. This means that a greater *per cent* error is tolerated by the simplified formula than by the original formula.

### Discussion

The findings do not relate to the degree of agreement between scores derived via the two formulas. They relate instead to whether or not the new formula is operationally a *simplified* version of the old or whether or not it is *simplified* in name only. The results suggest that the revised formula is superior with respect both to time taken and accuracy with which it is applied. The use, in this study, of untrained subjects shows that this superiority does not depend on training or previous experience but resides instead in the different method of counting required by the new formula. This formula, therefore, appears to be operationally a simplified version of the old one.

### Summary

A factorial experiment was undertaken to study the effects of various factors on the accuracy and time taken by naive subjects to perform readability counts. The factors investigated were: (1) difficulty of reading ma-

terial; (2) the type of count performed; (3) reading ability of persons performing the counts; and (4) sex.

The major finding was that the counting of one syllable words could be done in about three-fourths the time required for counting syllables. Boys performed the former count more accurately than the syllable count. This difference was not statistically significant among the girls.

A significant interaction effect was found between difficulty level and type of count. The syllable count was performed more accurately for easy material; the one syllable word count was performed more accurately for difficult material. Neither accuracy nor time taken was significantly associated with reading ability or sex.

It has been concluded that the new F, J, and P formula is truly *simplified* since it can be applied with a greater degree of accuracy and requires less counting time.

Received February 24, 1953.

Early publication.

### References

1. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, 3, 22-38.
2. Cochran, W. G., and Cox, G. M. *Experimental designs*. New York: Wiley, 1950.
3. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
4. Farr, J. N., Jenkins, J. J., and Paterson, D. G. Simplification of Flesch reading ease formula. *J. appl. Psychol.*, 1951, 35, 333-337.
5. Farr, J. N., Jenkins, J. J., Paterson, D. G., and England, G. W. Reply to Klare and Flesch re "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 55-57.
6. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
7. Flesch, R. Reply to "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 54-55.
8. Klare, G. R. A note on "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 53.
9. Nelson, C. W. *Use of factorial design in industrial relations research*. Research and Technical Report 6, University of Minnesota Industrial Relations Center. Dubuque, Iowa: Wm. C. Brown Company, 1950. Pp. 52.

## Reliability of the Original and the Simplified Flesch Reading Ease Formulas

George W. England, Margaret Thomas, and Donald G. Paterson

*University of Minnesota\**

Both Klare (7) and Flesch (5) in their attacks on the Farr, Jenkins, and Paterson (2) simplification of the Flesch reading ease formula (4) suggested that the reliability of the F, J, and P simplification formula would be lowered. Klare asserted that the simpler method would magnify each counting error and thus decrease reliability. Flesch attacked the idea that  $r$  between the original and the simplified reading ease scores would be higher for more *heterogeneous* materials<sup>1</sup> than for the employee handbooks used in developing the F, J, and P simplified formula and, in effect, implied that the reliability of the Flesch formula is impaired by the F, J, and P simplified formula.

Farr, Jenkins, Paterson, and England (3) were able to answer the Klare and Flesch criticisms with respect to the relative knowledge of syllabification required for both formulas and also with respect to the speed with which the new, simplified formula can be applied. A mean time of 82 seconds versus 147 or a saving of 65 seconds per 100-word sample was found. Discussion of the problem of reliability, however, was necessarily postponed with the following statement, "A thoroughgoing study of the reliability of both methods would be needed to settle this issue" (3, p. 56).

The present paper reports on this aspect of the problem.

\* England was research assistant in the Industrial Relations Center, Miss Thomas was a graduate student in psychology, and Paterson was professor of psychology and member of the staff of the Industrial Relations Center at the time the study was made. England is now with Personnel Research Staff of RCA at Camden, N. J., and the other two authors continue in their same roles at the University of Minnesota.

<sup>1</sup> The idea that  $r$  would be lowered if more *heterogeneous* materials were used is naive, statistically speaking.

### Procedure

*Data from House Organs.* During the spring quarter of 1952, 13 pairs of analysts computed reading ease scores by both formulas for each of 196 hundred-word samples drawn from 49 house publications.<sup>2</sup> Most of these analysts had participated during the winter quarter of 1952 in the prior study of the time required to compute reading ease scores by the Flesch method and by the F, J, and P simplified method. Stress was now placed on *accuracy* of counting syllables, one-syllable words, and sentence length as well as in the use of the Farr and Jenkins table (1) and the Farr, Jenkins, and Paterson table (2). One member of each pair used the old formula and the new formula in analyzing a given hundred-word sample and the other member of each pair did likewise for the same hundred-word sample. The 14 analysts formed 13 pairs and each pair, on the average, analyzed about 15 hundred-word samples. It is recognized that this procedure would produce *lower* reliability coefficients than would have been the case if one pair of experienced analysts had analyzed all 196 samples. It was anticipated, however, that the emphasis on accuracy of *all* the operations would tend to produce acceptable reliability data.

*Data from Books.* One analyst<sup>3</sup> undertook to compute reading ease scores by both

<sup>2</sup> Graduate students in Mr. Paterson's Seminar in Applied Psychology participated in the study. The work was done under the immediate supervision of George W. England who also assumed responsibility for the preparation of the statistical constants and for the reliability coefficients. The writers are grateful to the following students: Robert C. Becker, Sarah Ruth Cook, Ellen A. Corcoran, George W. England, Benno G. Fricke, Richard S. Hatch, Sulo N. Havu, Benjamin Lasoff, Raymond C. Lee, Jr., Paul W. Maloney, Ernest L. McCollum, Arthur C. McKinney, Charles Newstrom, and Margaret Thomas. Margaret Thomas conducted this phase of the investigation.

Table 1

Means, Standard Deviations and Reliability Coefficients for Analyst to Analyst Study of the Flesch and the Farr, Jenkins, and Paterson Simplified Reading Ease Formulas Applied to House Organs

Note: N = 196 hundred-word samples drawn from 49 House Organs with counts and computations made by 13 pairs of analysts.

	Analyst 1		Analyst 2		$r$ (Analyst 1 versus Analyst 2)
	Mean	S.D.	Mean	S.D.	
Sentence Length	20.3	7.0	20.3	7.1	.90
Syllable Length	159.6	15.3	158.7	15.6	.97
No. of One-Syllable Words	62.4	7.6	62.4	7.9	.95
Flesch R. E. Score	51.2	16.1	51.7	15.9	.96
F, J and P R. E. Score	48.0	14.1	47.8	14.1	.93

formulas for each of 196 hundred-word samples drawn from 28 books. Then, at a later date, this same analyst recomputed the data for 77 of the 196 samples drawn from 11 books. In this way, a basis was provided for computing test-retest reliability coefficients for the 77 samples.

### Results

*Data from House Organs.* The statistical constants and the reliability coefficients<sup>4</sup> are presented in Table 1. It will be noted that the means and sigmas obtained by each of a pair of analysts are quite close. The reliability coefficients shown in column 4 of Table 1 are all .90 or higher. As was true in the Hayes, Jenkins, and Walker study (6), the reliability of computing *average sentence lengths per hundred-word sample* is lower than for making the syllable counts. Furthermore, the evidence shows that *total syllable counts* and *counting the number of one syllable words per hundred-word samples* are made with a gratifyingly high degree of reliability (.97 and .95 respectively). The reliability coefficients of reading ease scores per hundred-word sample whether computed by the original Flesch method or by the simplified F, J, and P method are likewise high (.96 and .93 respectively). These reliability coefficients compare favorably with those re-

ported by Hayes, Jenkins, and Walker (6) for the original Flesch formula. Thus, no real loss in reliability has arisen by the introduction and use of the F, J, and P simplified formula.

*Data from Books.* The statistical constants and the reliability coefficients for a single analyst are presented in Table 2. Again, the means and sigmas of the first count or computation (test) and of the second count or computation (retest) by this analyst are quite close. Of more importance, however, is the fact that these hundred-word samples drawn from 11 books represent far more *heterogeneous* materials than was true of the samples drawn from the house organs. The sigmas in Table 2 when compared with the sigmas in Table 1 clearly prove this point. The range of the original Flesch reading ease scores for these 11 books was from 100 for "*Fun with Dick and Jane*" to 26 for "*Personality, a Psychological Interpretation*." As a matter of fact, the 11 books were selected from *all* the difficulty levels. And, as would be expected, the test-retest reliability coefficients are much higher. In fact, they approach unity. This is due to the combined operation of the greater *heterogeneity* of materials sampled and having only a single "compulsive" analyst make all counts and computations. The results closely approximate the high analyst to analyst reliability coefficients reported by Hayes, Jenkins, and Walker (6).

<sup>4</sup> The reliability coefficients in Table 1 may be thought of as "alternate form reliability coefficients" since they are "analyst to analyst" reliability coefficients.

Table 2

Means, Standard Deviations and Test-Retest Reliability Coefficients for a Single Analyst Study of the Flesch and the Farr, Jenkins, and Paterson Simplified Reading Ease Formulas Applied to Books

Note: N = 77 hundred-word samples drawn from 11 books with all counts and computations made by a single analyst.

	First Count or Computation (Test)		Second Count or Computation (Retest)		Test Retest <i>r</i>
	Mean	S.D.	Mean	S.D.	
Sentence Length	19.8	10.6	20.0	10.7	.95
Syllable Length	146.6	19.7	146.7	19.6	.99
No. of One-Syllable Words	69.6	8.5	69.4	8.7	.99
Flesch R. E. Score	62.4	24.7	61.7	24.4	.99
F, J, and P R. E. Score	59.0	21.6	58.7	21.7	.97

### Intercorrelations between Original and Simplified Formulas

Two intercorrelations between the original and simplified RE scores for 196 samples from 49 house publications were computed: (a) for analyst 1,  $r$  was +.84; and (b) for analyst 2,  $r$  was +.87. The intercorrelation between the original and simplified RE scores for 196 samples from 28 books for a single analyst was +.94 and for the 77 samples from 11 books for the same single analyst,  $r$  was .97. The intercorrelation for the original and simplified RE scores for the *averages* of the 28 books (7 one-hundred word samples each) was +.97. Thus, the original and the simplified RE scores are comparable when computed by a single, fairly experienced and compulsive analyst.

### Summary

The reliability of the original and the simplified Flesch reading ease formula based on (a) samples drawn from house organs, using 13 pairs of relatively inexperienced analysts; and (b) samples drawn from books, using a single, more experienced analyst is reported. The findings confirm the earlier reliability study by Hayes, Jenkins, and Walker (6) and show that both the original and the sim-

plified Flesch reading ease formulas are highly reliable. With *heterogeneous* materials and a single "compulsive" analyst, test-retest reliability coefficients from +.95 to +.99 were obtained. Intercorrelations between the original and simplified formulas are likewise "high."

Received February 10, 1953.

Early publication.

### References

1. Farr, J. N., and Jenkins, J. J. Tables for use with the Flesch readability formulas. *J. appl. Psychol.*, 1949, 33, 275-278.
2. Farr, J. N., Jenkins, J. J., and Paterson, D. G. Simplification of Flesch reading ease formula. *J. appl. Psychol.*, 1951, 35, 333-337.
3. Farr, J. N., Jenkins, J. J., Paterson, D. G., and England, G. W. Reply to Klare and Flesch re "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 55-57.
4. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
5. Flesch, R. Reply to "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 54-55.
6. Hayes, Patricia M., Jenkins, J. J., and Walker, B. J. Reliability of the Flesch readability formulas. *J. appl. Psychol.*, 1950, 34, 22-26.
7. Klare, G. R. A note on "Simplification of Flesch reading ease formula." *J. appl. Psychol.*, 1952, 36, 53.

## Validity of Readability Formulas \*

Charles E. Swanson

*Institute of Communications Research,  
University of Illinois*

and

Harland G. Fox

*Industrial Relations Center,  
University of Minnesota*

Whether readability formulas can be used to predict more or less success in all printed communication is not known. Some authors suggest their formulas will discriminate among articles expected to get more or less readership, understanding, etc. Dale and Chall (1) use the title, *A formula for predicting readability*. Flesch (2) says more readable writing will "appeal to readers" and cites an experiment by Swanson (9) where readership was the criterion. However, the excellent bibliographies of Hotchkiss and Paterson (5), Flesch (2) and Klare (6) show few validation studies using comprehension and retention as criteria.

In their pioneering study Gray and Leary (4) found 24 factors of style related to reading comprehension of adults. Gray and Leary, Dale and Chall, and Flesch reduced these to a few factors. Their findings agreed on word difficulty and sentence length. Flesch also used personal references in one of his formulas.

In two experiments with articles in a mid-western farm paper Ludwig (7) varied one factor at a time, word difficulty and personal references. His test articles were each read by more than 40 per cent of the two samples of farmers. Readership differences between the experimental pairs of articles were small and were not significant.

Analysis of Ludwig's findings suggested several hypotheses:

Readability factors would have maximum effect when two or more positively related

factors were varied. Easier words and shorter sentences, for example, should result in increases of comprehension, other things being equal.

Where more than 40 per cent of an audience selects and reads an article, less gains in effect can be expected from improved readability. Also, where lesser proportions of an audience read an article, the more that gains may come from increases in readability.

Motivational factors inherent in content, such as subject matter, probably are more important, generally, than readability where individuals select what they want to read and learn from printed media. For example, comic strips are easy to read but vary widely in readership, or audience interest. One comic strip may reach 70 per cent and another strip in the same day's newspaper reach 20 per cent of the same audience.

Readability factors might be more important than motivational factors where individuals are required to read and study and are tested on their learning. This would be the case in classroom and training situations.

### The Present Experiment

In this study easier and harder versions of 12 articles were published in three issues of a paper sent monthly to employees of a mid-western company. Four articles appeared each month. The 296 employees were randomized into two groups, "easy sample" and "difficult sample."

Easy sample received copies of the newspaper with easier versions of the 12 articles. Difficult sample received the same newspaper with the harder versions.

The 12 articles concerned company products, company history, safety program, and the working agreement which covered wages, hours, and working conditions.

Effects of the versions were determined by these criteria: (1) Retention; measured by a 43-item test of multiple-choice questions

\* Grateful acknowledgment is made to the Graduate School, University of Minnesota, for the research grant to finance preliminary analysis, field work, and part of analysis of results. Intensive analysis of content and information tests were supported under Contract N60NR-246, T.O. 4, Office of Naval Research, with the senior author as responsible investigator. Aid was provided by staffs of the Industrial Relations Center and Research Division, School of Journalism, University of Minnesota, and by Drs. James J. Jenkins and Robert L. Jones. The senior author is indebted to Dr. George M. Klare, University of Illinois, for his critique of the analysis.

based on the 12 articles; (2) Readership; measured on easier or harder versions of two articles; and (3) Comprehension; measured by a 10-item test given before and after exposure to easier or harder versions of two articles.

Four other instruments were used. They involved general opinions about company and union, general satisfaction with one's job (11), Sanford's authoritarian-equalitarian scale (8), and Goossen's disguised intelligence test (3).

Each subject followed this sequence in an interview: (1) Took the 43-item information test; (2) Read easier or harder versions of two articles; (3) Reported whether he had read the two articles when they appeared in the company newspaper. (Sixty per cent had read the articles. Actually these subjects were reading the articles a second time in the comprehension test.); (4) Took 10-item information test on the two articles. (The 10 items were included in the 43-item test.); and (5) Answered four questionnaires on general opinions about company, union and job, authoritarian-equalitarian personality aspects, and intellectual ability.

#### Readability Differences

Three questions concern changes from harder to easier versions. What were the differences in readability? Could some factors decrease comprehension and so decrease positive effects of other factors? Did easier and harder versions have the same information content?

The following readability differences appeared.

*Formula scores.* By the Flesch formula, the easier versions had a mean score of 73 (fairly easy) whereas the harder versions scored an average of 59 (fairly difficult). The Dale-Chall formula gave similar results. The easier versions had a mean Dale-Chall score of 7th-8th grade compared with a score of 11th-12th grade for the harder versions.

*Number of words.* The easier versions had fewer words, an average of 284, while the harder versions had an average of 332 words. The easier versions totaled 3,410 words and the harder versions totaled 3,983 words.

*Flesch human interest index.* The easier versions had a mean score of 46 (very interesting) and the harder versions a mean of 17 (mildly interesting).

*Sentence length.* The easier versions had an average sentence length of 13.0 words. The harder versions had an average sentence length of 19.4 words.

*Syllables per 100 words.* The easier versions had 142 syllables per 100 words and the harder versions 161 syllables per 100 words.

*Unfamiliar words.* As scored by the Dale-Chall list of 3,000 unfamiliar words, the easier versions had 11.6 per cent unfamiliar words whereas the harder versions had 20 per cent unfamiliar words.

*Verbs and adjectives.* The easier versions had 130 verbs per 100 adjectives. The harder versions had 89 verbs per 100 adjectives.

This study could not answer the question of whether some of these or other readability factors cancelled out comprehension gains. No previous research had been published at the time of designing the study to suggest this possibility. The Gray-Leary and Swanson investigations indicated that readability factors such as these would combine for positive effects.

In the opinion of three judges the information content of the easier and harder versions was the same. They used the multiple-choice questions as aids to their judgments. No method was known to the investigators by which information content could be classified and its similarity between easier and harder versions defined quantitatively.

Differences in effects of easier or harder versions could not be attributed to subject matter.

Whether differences could be attributed to fewer words used in easier versions is a question of whether details were amplified in the harder and longer versions. Wilson (10) used versions 300, 600 and 1,200 words in length. She found that amplification was helpful only where the reader had difficulty with concepts. Any advantage in this respect might be in favor of the longer versions. However, the investigators believed that the information content and amount of amplification were held constant. Again, no quan-

titative method was devised to permit other investigators to check this point.

### Characteristics of the Two Samples

The two samples were interviewed under the same conditions by the same group of interviewers in the company's dining hall. A total of 130 interviews was completed (67 easy sample and 63 difficult sample).

Attrition of the original population of 296 employees was due to several factors. A total of 96 were "laid off"; 6 quit and the remainder were on night shift or vacation or were ill or were illiterate. No significant differences between the samples could be attributed to these factors.

The two samples did not differ significantly on the following social characteristics:

*Social and individual.* Age, sex, years of schooling, marital status, mean scores on the authoritarian-equalitarian scale, and the Goossen disguised intelligence test.

*Job and union.* Years with the company, years in current job, union membership, years in the union, readership of a union paper, and opinions about company, union, and job.

Easy sample had more employees high and low in intellectual ability as measured by the Goossen disguised intelligence test. Mean test scores, however, did not differ significantly.

Compared with the general population of American adults, these two samples of 128 employees included more females (60 per cent); younger persons (36 per cent from 20 to 30); more schooling (55 per cent with some high school and 13 per cent with some college).

More than 60 per cent had worked more than five years for this firm and 65 per cent were union members. Of those who were union members two-thirds had been union members more than five years.

### Results

*Retention.* The two samples did not differ significantly in mean scores on the 43-item test based on information in both versions of the 12 articles.

Item analysis showed the two samples did not differ significantly on 37 of the 43 in-

formation questions. Of six items where differences were significant, easy sample had higher scores on two and difficult sample had higher scores on four questions.

No consistent patterns appeared in kinds of items on which one sample succeeded more than the other. Easy sample had higher scores on questions about annual sick leave and a provision of the working agreement. Difficult sample had more success on two items about company history, the "cause" of hard water, and the name of an official who bargained with the union.

The two samples did not differ in remembering information from easier or harder articles. The formulas did not appear to measure factors in the articles related to differences in retention.

*Readership.* The two samples did not differ significantly in readership. Of easy sample, 65 per cent ( $n = 67$ ) read both articles; of difficult sample 61 per cent ( $n = 63$ ) read both articles.

The easier versions did not reduce the proportion who failed to read both articles. Of easy sample 22 per cent did not read either article and 29 per cent of difficult sample did not read either article.

Neither the readability formulas nor the Flesch human interest index seemed to measure factors in the articles related to differences in readership.

*Comprehension.* Easier and harder versions of the two articles used in the test of readership also were used to test comprehension. Subjects read easy or difficult versions of the two articles in the test situation; immediately after reading they answered 10 questions based on the two articles. These 10 questions had been included in the initial 43-item test. Mean scores on this 10-item test, before and after reading in the test situation, are shown in Table 1.

Easy sample did significantly better than difficult sample on the 10-item after-reading test. However, the two samples did not differ significantly on the before-reading test.

This result indicates that the readability formulas did measure factors in the articles which related to differences in comprehension.

An analysis of the 10 items showed that

Table 1

Mean and Variance Significance Tests for Sample Easy and Sample Difficult on Information Tests  
 Note: Sample Easy  $N = 67$ ; Sample Difficult  $N = 63$ .

Variables	Mean Sample Easy $n = 67$	Mean Sample Difficult $n = 63$	S.D. Sample Easy	S.D. Sample Difficult	$F$	$t$
43-Item Information Test	20.93	22.29	5.96	5.54	1.16	1.34
10-Item Test Before Reading	5.25	4.87	1.82	1.78	1.04	1.37
10-Item Test After Reading	8.03	7.16	1.91	1.85	1.06	2.61**
Gains in Correct Response on 10-Item Test After Reading	2.78	2.30	1.72	1.77	1.06	1.54

\*\* Significant at the 1% level.

easy sample made consistent gains in comprehension over difficult sample. None of the gains appeared important except for one item. On this question easy sample showed four times as high a gain (54 per cent) in correct responses as difficult sample (13 per cent).

The evidence, both qualitative and quantitative, showed that readability indices could be used to predict differences in comprehension between two versions of the same material.

*Readers vs. non-readers.* In the reader-ship test of two articles, about two-thirds of the 128 employees read both articles. The remainder either ignored both items or read one. Would readers have higher information scores than non-readers? Obviously, much information could have been learned from personal experience or other sources. Yet one might expect readers to know more; the reading behavior might be symptomatic of efforts to learn similar information from other sources.

By the reading criterion, subjects were divided into three groups: 80 who had read both test articles in the company newspaper; 15 who had read one; 33 who had read neither. The two extreme groups, readers and non-readers, were compared.

On the 43-item information test the readers had a mean of 23.5 items, or 55 per cent, correct. Non-readers had a mean of 18.5, or 43 per cent, correct. This difference was highly significant ( $t = 5.11$ ).

Item analysis (by reader and non-reader)

showed 11 significant differences. In each case readers were more successful.

Of the remaining 32 items readers had a higher proportion of correct response on 29 items. By the sign test, this was a highly significant difference.

Readers had significantly higher mean scores than non-readers on the 10-item test before but not after reading the two test articles. The non-readers gained more in comprehension. From *before* to *after* reading, the non-readers gained on the 10 items an average of 2.9 items correct, compared with 2.2 for readers. This was a significant difference ( $t = 2.00$ ).

Whether readers had more intellectual ability than non-readers became an important question. They did not differ in years of schooling or in Goossen disguised intelligence test scores. This suggested that readers might differ from non-readers on other social characteristics which could explain differences in motivation, or interest in the material.

Ten factors were analyzed for clues to differences in motivation between readers and non-readers. These were age, sex, years with the company, years on the specific job, union membership, years in the union, readership of a union paper, general opinions about company and job, authoritarian-equalitarian score, and union activity. Readers and non-readers did not differ on these factors. No characteristic discriminated between those employees more and less motivated to read and learn information from the company newspaper.

Table 2

Mean and Variance Significance Tests for Readers and Non-Readers on Information Tests  
 Note: Readers  $N = 80$ ; Non-Readers  $N = 33$ .

Variables	Mean Readers	Mean Non- Readers	S.D. Readers	S.D. Non- Readers	F	t
43-Item Information Test	23.46	18.50	4.83	5.98	1.54	5.11**
10-Item Test Before Reading	5.56	4.36	1.66	1.74	1.12	3.64**
10-Item Test After Reading	7.80	7.30	1.75	2.21	1.61	1.26
Gains in Correct Response on 10-Item Test After Reading	2.24	2.94	1.61	1.86	1.36	2.00*

\* Significant at the 5% level.

\*\* Significant at the 1% level.

### Summary and Discussion

When easier and harder versions of 12 articles were printed in three monthly issues of a company newspaper and two samples of 128 employees were tested, it was found:

1. Subjects exposed to harder versions succeeded as well on a 43-item information test as those exposed to easier versions.

2. Harder versions succeeded as well as easier versions in attracting readers to two articles.

3. Subjects who read easier versions of two articles in a test situation did significantly better on a 10-item test of comprehension than those who read harder versions.

This result indicates that readability formulas can predict some differences in comprehension between versions of the same material.

4. Readers of two articles were more successful on the 43-item test of information in the 12 articles than those who had not read either of the two articles tested for comprehensibility.

These results indicate that readability formulas can be used to predict differences in comprehension between two versions of the same material. However, the findings do not support the utility of such formulas in predicting differences in readership, and retention for similar material, conditions, and time periods. Even combined treatment of readability factors, such as was attempted in this study, did not influence retention.

One factor limiting these results is the relatively high interest (readership by 60 per cent of the samples) in two of the articles.

The lack of differences in retention be-

tween easier and harder versions suggests that investigation of motivational factors inherent in content is most crucial where individuals select what they want to read and learn. This does not gainsay the possibly greater importance of readability where individuals are required to read and study as in classroom and training situations.

Received April 4, 1952.

### References

1. Dale, E., and Chall, J. S. A formula for predicting readability. *Educ. Res. Bull.*, Ohio State University, 1948, 27, 11-20 and 37-54.
2. Flesch, R. *How to test readability*. New York: Harper & Brothers, 1951.
3. Goossen, C. F. *The construction and validation of a disguised intelligence test to be used in public opinion interviewing*. Unpublished Ph.D. thesis, University of Minnesota, 1949.
4. Gray, W. S., and Leary, B. E. *What makes a book readable*. Chicago: University of Chicago Press, 1935.
5. Hotchkiss, S. N., and Paterson, D. G. Flesch readability reading list. *Personnel Psychol.*, 1950, 3, 327-344.
6. Klare, G. R. *Evaluation of quantitative indices of comprehensibility in written communication*. Unpublished Ph.D. thesis, University of Minnesota, 1950.
7. Ludwig, M. Hard words and human interest: their effects on readership. *Journ. Quart.*, 1949, 26, 167-171.
8. Sanford, F. H. *Authoritarianism and leadership*. Philadelphia: Stephenson-Brothers, 1950.
9. Swanson, C. E. Readability and readership: a controlled experiment. *Journ. Quart.*, 1948, 25, 339-343.
10. Wilson, M. C. The effect of amplifying material upon comprehension. *J. educ. Psychol.*, 1947, 38, 149-156.
11. Yoder, D., Heneman, H. G., Jr., and Cheit, E. F. *Triple audit of industrial relations*. Bull. 10, Industrial Relations Center, University of Minnesota, 1951.

## A Note on Pre-testing Public Opinion Questions

Robert C. Nuckols

*Life Insurance Agency Management Association, Hartford, Conn.*

There are probably no individuals engaged in measuring attitudes or public opinion who would not agree that it is wise to pre-test questionnaires. Many would probably say that the conventional pre-tests are conducted efficiently and result in well designed and adequately worded questionnaires. About this latter point there is some doubt.

Several years ago this writer conducted a pilot study of respondent comprehension using a battery of "typical" opinion questions. The results of this study seem to shed some light on the question of the adequacy of our present pre-testing methods.

### Procedure

Nine questions were chosen from "The Quarter's Polls," and were presented to a randomly selected group of 48 middle-income respondents in Cincinnati and Centerville, Ohio. The questions were selected to cover a wide range of reading difficulty as judged by the Flesch readability formula.<sup>1</sup> Of the nine questions, one had a difficulty equal to the adult average level as defined by Flesch, four were above and four below this level of difficulty. A second criterion for the selection of questions was that they be of topical interest to the respondents at the time this study was being conducted.

To test the respondents' comprehension of the questions, a rather simple procedure was used. The question was presented to the respondent and after his answer had been given he was asked to repeat in his own words the meaning of the question as nearly as he could. The interviewer then recorded the respondent's interpretation verbatim. The order of question presentation was varied from respondent to respondent.

There are a number of criticisms that one could level against this method of measuring comprehension. It may be argued that

merely because a person can parrot a question, it does not necessarily follow that he comprehends its meaning. On the other hand, if a respondent gives a faulty interpretation, it seems fairly safe to conclude that he did misinterpret it. This would probably lead to an underestimation of comprehension, certainly not an overestimation.

The respondents' interpretation of each of the questions was judged to fall into one of four categories: (a) correct interpretations, leaving out no vital parts; (b) generally correct replies, or replies in which no more than one of the parts was altered or omitted; (c) partially wrong interpretations, but showing the respondent knew the general subject of the question; (d) completely wrong interpretations or no-response. As an example of the scoring take the question: "Suppose the government had no control over how the businesses are run in this country, who do you think this would help the most—the people as a whole, or those who run big businesses, or those who run small businesses?"

A partially correct interpretation was: "If there weren't any control, which would have the greater power—the small business or the larger."

A partially wrong interpretation was: "'Bout government owning business—who would benefit most, big businesses or small businesses."

Or: "Just who would get the business—the big guy or the little guy?"

An example of a completely wrong interpretation was: "Something about having a President. If he does things that people don't agree with, they have a right to tell him—like Walter Winchell."

The responses to the questions were judged individually by each of two judges. In case of disagreement, the response was discussed until agreement could be reached as to which interpretation category it belonged.

<sup>1</sup> Flesch, R. *The art of plain talk*. New York: Harper and Brothers, 1946.

## Results

There were 430 question interpretations of which 73, or 17.0 per cent, were either wholly or partially wrong. Two respondents did not make an interpretation of one question because, in one case, the telephone rang and, in the other case, something was boiling over on the stove.

The findings would not be startling if one could say that these questions have now been pre-tested and can be re-worded so as to make them more comprehensible. However, the questions used in this study had already been presented to large cross sections of the general public by well known polling organizations. That is, these are questions *after* they presumably have been subject to the usual pre-test.

If the questions had been asked of the respondents and only their answers recorded in the usual way these errors of comprehension would not have been detected. *In no instance did a respondent say that he did not hear a question or that he misunderstood it.* The questions were asked, answers given, and all seemed well.

If one grants that some degree of respondent comprehension may be missed in the usual pre-test, it still may be asked if this error contributes to any inaccuracy in poll results. From this study, the answer seems fairly clear. Four questions contributed by far the most to the total amount of miscomprehension. In two of these questions, there was a marked and statistically significant tendency for those not comprehending to reply "don't know." On one question there was a significant tendency for the non-comprehending respondent to answer "approve" to a question dealing with the United Nations. This institution was enjoying a high degree of popularity at the time this study was conducted, and hence likely to elicit a favorable stereotype from a respondent hearing the words "United Nations" but miscomprehending the intent of the question. In this instance, the question was inquiring about placing atomic energy under UN control. There was no tendency evident for

Table 1

Relation Between Readability and Comprehensibility of Nine Opinion Poll Questions

Estimated Reading Grade Placement (Flesch Score)	No. of Miscompre- hensions
5.8	1
6.1	1
7.2	14
7.6	14
8.5	14
11.0	1
12.8	14
14.0	3
17.2	11

those miscomprehending the remaining question to reply differently from the rest of the sample. This in itself might be damning to that question.

Because the sample of questions is small, and because several of the questions received equal miscomprehension scores, the correlation between comprehension and readability has not been presented here. This study was not designed to be a validation of the Flesch index; however, since there may be some interest in the relationship found, the Flesch score of each question and the number of persons miscomprehending the question are presented in Table 1. The number miscomprehending the questions included those making wrong and partially wrong interpretations.

## Summary

From the results noted here, it would seem that conventional pre-testing fails to uncover many questions that are later misinterpreted by respondents in the main survey. And it would seem that the failure to word some questions so as to bring respondent comprehension to a maximum may result in distortion of the survey results. Hence, a few extra minutes spent gaining some rough measure of comprehensibility of the questions may well pay ample dividends in increased survey accuracy.

Received May 14, 1952.

## A Study of Respondent Forewarning in Public Opinion Polls \*

Robert C. Nuckols

*Life Insurance Agency Management Association, Hartford, Conn.*

Most of us have had the experience of being called upon unexpectedly to give an opinion about some question, state a course of action, or criticize some proposal in an intelligent manner. It is possible that in such situations we made replies that we later recognized as missing the point, as not fully expressing our position, or that would have been more valuable if we could have thought of this, that, or the other alternative. It is conceivable that a large proportion of respondents to the typical opinion poll find themselves in a similar position. The respondent may give a forced answer to the persistent probing of the interviewer. However, after the interviewer has gone these respondents may recall many pertinent bits of information or opinion that would clarify, amplify, or even change their original position. These additional remarks on the part of the respondent should be of some interest in the analysis of opinion.

This study was undertaken to determine the effects of forewarning the respondents of a "typical" opinion poll of the purpose and nature of the approaching interview.

It is hypothesized that forewarning by means of an introductory letter will give the respondent an opportunity to think about and discuss the various topics listed in the letter and so be prepared to give more detailed and thought out answers than he would with no such opportunity. It is also hypothesized that by forewarning, the respondent will be more prepared to cooperate with the interviewer and therefore make the interview more enjoyable, both from the interviewer's and the respondent's point of view.

\* This study is part of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Pennsylvania State College. The study was completed while the investigator was a fellow in psychology of the Britt Foundation.

### Method

Two community surveys, one in Altoona and the second in Williamsport, Pennsylvania, were conducted to test these hypotheses. The samples were drawn from the most recent city directory on an every  $n$ th dwelling unit basis. One of these sub-samples, comprising 60 per cent of the total sample in each city, was designated for the sending of the forewarning letter. Letters were sent to more than one-half of each sample to allow for the normal number of substitutions and refusals. The letters were sent so as to be received at least three days before the interview.

The forewarning letter read as follows:

Dear Residents of Altoona:

Many things, both big and small, are important to you in deciding whether or not a city is a good place in which to live. In an effort to make Altoona a better place in which to live, a study is being made by Pennsylvania Surveys at the request of the Altoona Chamber of Commerce.

To meet this aim it is important for you, as residents of Altoona, to speak your thoughts and opinions on several topics of community interest. Only you and your neighbors can paint a true picture of your city. We feel sure that you will cooperate to help make this project a success.

On Tuesday, November 28, a representative of Pennsylvania Surveys will call on you at your home. He, or she, will ask:

About the transportation service, within Altoona, and into and out of Altoona.

About business and industry in Altoona.

About services provided by the city government of Altoona.

About the amount and kind of public recreation available in Altoona.

About the housing situation in Altoona.

About the public schools.

For over-all suggestions that would make Altoona a better place in which to live.

We realize that you are concerned with many problems other than those of a purely local interest. Therefore, we have sent you this letter so that when the representative calls you will have had some time to think about these problems. We hope you will think about these top-

non-forewarned. As is noted in the table, this hypothesis was not confirmed in either city.

In both cities there was only a negligible tendency for the forewarned respondents to give more answers to the open-end questions than the non-forewarned. In neither city was the difference close to being significant.

The forewarned group in the preceding analyses contained respondents who did not report receiving the letter, and those who reported not understanding the topics. To further test the effects of forewarning, those respondents in Williamsport who claimed to have received the letter and to have understood its meaning were selected out of the over-all forewarned group. A sample of respondents was drawn from the non-forewarned group to match as completely as possible the informed-forewarned in respect to age, sex, socio-economic status, and educational attainment. These two groups were then analyzed on the same variables as discussed above. Only those questions covering topics that were mentioned specifically in the forewarning letter were included for analysis.

If there was any marked tendency for these informed-forewarned respondents to change their responses on the basis of the opportunity to discuss and think about the topics then it should turn up in this matched sample analysis. However, in no instance were the hypotheses verified. That is, there was no tendency for the informed-forewarned group to give more responses to open-end questions, fewer "don't know" responses, less stereotyped or fewer non-reality replies, or to accept more extreme statements of opinion.

Returning once more to the full samples, the hypotheses concerning respondent cooperation were analyzed next. Because some of the interviewers failed to record refusals, it was impossible to determine the effects of forewarning upon the refusal rate. However, with data obtained from the original sample listings it was possible to analyze the rate of substitution at forewarned and non-forewarned addresses. A difference of six per cent was obtained in the direction of fewer substitutions at addresses to which let-

ters had been sent. This difference was significant at the 10 per cent level of confidence.

Interviewer ratings of the respondent's cooperativeness, eagerness to discuss the questions, and apparent information showed differences in favor of the hypothesis. In both Altoona and Williamsport the forewarned respondents were rated significantly more cooperative than the non-forewarned. In neither city was a significant difference found in respect to the respondent's eagerness to discuss the questions, but the differences that did exist were in the predicted direction. The interviewers did not rate the Altoona forewarned respondents as being more informed; however, there was a significant difference in favor of the forewarned group in Williamsport.

### Discussion

When considering the questions individually it was found that the number of significant differences in this study could have been found on the basis of chance alone. Secondly, the differences that were found were not consistently in the predicted direction, nor were they consistent between the two cities. The matched-sample study, which tested the hypotheses under the most rigorous conditions, did not disclose any differences that would uphold the hypothesis that the forewarning letters lead to more meaningful or a greater number of responses.

The general lack of effectiveness of the forewarning letter may have resulted from several uncontrolled variables. It is possible that the forewarned respondents, if probed more completely or given more time to give their answers, would have given more responses and responses with more meaning. It is also possible that the older and stronger opinions, those most likely to be stereotyped, would be given first. If the interviewers did not record all the answers or were in a hurry to get to the next question the effects of forewarning might be nullified. However, while this may all be true, it is held that the interviewers were well motivated, did a competent job, and were comparable in ability to the typical interviewer used in many market research studies.

Another factor that might account for the negative findings is the letter. If the letter had not been mimeographed, but rather had been made more attractive or had spelled out the topics in a simpler or more understandable fashion, the respondents might have been more motivated to read the letter and take the suggested action. On the other hand, the letter had been pretested on a small sample in Altoona and checked for readability. Short sentences, short words, and large type were used, hence the letter should have been understandable to any person who could read a newspaper. While a more attractive letter might have secured more readers, the results of the matched-sample analysis showed readers to respond no differently from the non-forewarned.

The negative findings might have resulted because of the nature of the survey content. The respondents might have been more inclined to think about and discuss the national administration or the most recent baseball trades, rather than purely local issues. Here we may have the most logical explanation of the findings. This study may well have served to point out once again the public's indifference to civic affairs. In many of our cities well-documented exposés of civic maladministration or the pressing need for certain improvements fall on an unresponsive public. It may not be too far-fetched to believe that the forewarning letter met a similar fate.

The interviewer ratings are subject to the criticism that they were made after determining if the respondent had received a fore-

warning letter. Nevertheless, there is some subjective evidence that tends to uphold the general validity of the ratings.

The interviewers either volunteered or were asked whether or not they had the feeling that the forewarning letter made any difference in the respondents' cooperativeness. Many of the interviewers reported that they felt the letter did help in securing rapport and no interviewer reported that the letter made the respondent more suspicious or uncooperative. The interviewers claimed that they could predict the forewarned respondent with some accuracy before asking for knowledge of the survey. Moreover, the interviewers were not told the purpose of this study. They knew that some respondents would know of their coming; however, they were led to believe that this was primarily a check on their honesty in meeting the assignment. Therefore the hypothesis of increased cooperation would have to be arrived at individually during the course of the interviewing period.

If neither of these lines of argument validates the assumption of increased respondent cooperation it might be further argued that the question of the validity of the ratings is unimportant. If interviewers believe that forewarned respondents are more cooperative it makes very little difference whether they truly are more cooperative or not. It would seem that forewarning by mail can be an effective factor in making interviewing a more pleasant occupation, and that it can be done fairly inexpensively.

*Received June 16, 1952.*

## Influence of Ink Color on Handwriting of Normal and Psychiatric Groups\*

Walter A. Woods

*Richmond Professional Institute, Richmond, Virginia*

Color psychotherapy has aroused new interest in recent years and efforts have been made to re-establish the reputation of this declining field. Some writers have suggested that color vision is influenced by emotional states. Kravkov (3) has found that, under adrenergic influence, the retina is more sensitive to blue-green and less sensitive to red-orange.

From such studies it has been inferred by some that colors, as sensations (apart from symbolic content), are influential in producing states of emotion. In a recent popular book, *Color psychology and color therapy*, Birren (1, p. 150), draws upon such research to conclude: "To state a principle, it seems that the immediate action of any color stimulation is followed in time by a reverse effect. Red increases blood pressure, which later becomes normally depressed. Green and blue decrease blood pressure and later cause it to rise. . . ."

Birren relies on the work of Goldstein (2) for this generalization. He calls attention to Goldstein's observation (1, p. 149): "One could say red is inciting to activity and favorable for emotionally-determined actions; green creates the condition for meditation and exact fulfillment of the task. Red may be suited to produce the emotional background out of which ideas and actions will emerge; in green these ideas will be developed and the actions executed."

One might inquire concerning the basis on which Goldstein formulates these "principles." A report of his research is found in an article appearing in *Occupational Therapy and Rehabilitation* (2). Inasmuch as he merely refers to the research and describes neither

procedure nor data, it is impossible to determine how he arrived at his conclusions. The general nature of his findings are: any activity which takes place under red light or in which red equipment is used will tend to be performed in a more emotional manner whereas activity engaged in under green light or with green equipment will be "thoughtful" in nature. He describes an experiment (no data included) in which it was demonstrated that a subject with arms extended in front would, when illuminated with red light, tend to move his arms outward. If illuminated with green light, he would tend to draw the arms together in front of the body. He also discusses the influence of colored light or colored ink on handwriting. He found: "Words written in red ink or green ink (if the patient pays attention to the color) show different size of letter and different distances between the letters. Handwriting in green light or with green ink is much more similar to the normal handwriting than that in red light or in red ink" (2, p. 149).

Contrary findings are reported by Vollmer (9). He is unable to verify that arms of the subjects held forward and parallel deviate toward red and away from blue light.

Lukens and Sherman (4) found that the use of red, black or white materials by patients in weaving produced no differential results in woven objects.

In view of the inconclusive and conflicting nature of the evidence on which much of the contemporary opinion concerning color therapy is based, fundamental, planned experiments are necessary. The present article reports such an experiment.

### The Experiment

A total of 132 subjects were used. Of these, 66 were college students and 66 were patients in the State Mental Hospital at

\*From research conducted at Fort Hays Kansas State College. The author is indebted to Dr. J. T. Naramore, Supt., and Alexander J. Robinson, Clinical Psychologist, Larned State Hospital, for their help in this project.

Larned, Kansas, all classified as psychotic, psychoneurotic, or psychopathic personality and all in a state of remission suitable for occupational therapy and engaged in occupational therapy programs.

Each subject was asked to write the following statement in each of three colored inks, red, green, and black, and with a penholder which corresponded to the ink color.

Dear Joe, We received your letter and expect to see you next week, (signature)

Subjects were asked to write this statement on a sheet of white paper,  $5\frac{1}{2} \times 8\frac{1}{2}$  inches. Their attention was repeatedly directed to the fact that different ink colors were being used.

The particular statement was selected after preliminary experimentation, since it met the following requirements: (1) it was of such length that the average writer would not be tempted to cram it onto one line but could easily write it on two lines (length of the material should not influence choice of size or form of handwriting); (2) it was not so long that fatigue would be introduced; and (3) it was symbolically as meaningless as possible yet retained literary form.

The two major groups were subdivided into six sub-groups of 11 normals and 11 abnormals, in order to equalize the effect of the order in which the different colors were used. Group I used inks in the order R B I G; Group II, R G B I; Group III, G R B I; Group IV, G B I R; Group V, B I R G; and Group VI, B I G R.

This design supplied a total of 396 handwriting samples, 132 in red ink, 132 in black ink and 132 in green ink, one third of each ink color having been written first, one third second and one third last in the series.

Handwriting samples were measured on a millimeter scale, and means were determined for each sub-sample. Variance estimates of the sub-groups and major groups were made.

### Results

Sub-sample, border, and total sample means are given in Table 1. Color (column) means do not differ appreciably, nor do the order of writing (row) means. However, means for

Table 1

Means of 18 Groups of 22 Samples Each ( $nk$  equals 396) of Handwriting Classified According to: (1) Ink Color in Which Written; (2) Order in Which Ink is Used; and (3) Psychiatric Classification of Writer (Measurements in millimeters)

	NP Classification	Color of Ink			Order Means
		Red	Green	Black	
Order of writing	1 Normal NP	20.9	22.4	21.7	21.7
		26.9	26.5	27.1	26.9
	2 Normal NP	21.6	21.1	21.5	21.4
		27.4	27.1	24.7	26.4
	3 Normal NP	22.2	21.7	22.1	22.0
		26.4	26.7	27.5	26.8
Color means	Normal	21.6	21.8	21.7	21.7
	NP	26.9	26.8	26.4	26.7

### Analysis of variance

Source	df	Variance Estimate
Ink Color	2	.7
Order of Writing	2	11.26
NP Classification	1	3517.76
Color/order	4	14.22
Color/NP Class	2	3.53
Order/NP Class	2	.52
Order/Color/NP	4	18.20
Individual Diff.	359	9.66

$$\frac{\text{Order}}{\text{Ind. Diff.}} = 1.1; \frac{\text{NP Class}}{\text{Ind. Diff.}} = 364.1; \frac{\text{Color}}{\text{Ind. Diff.}} = 1.4$$

normal and for psychiatric groups differ in every instance.

$F$  test reveals that variance ratios in every instance are such that they would be expected by chance, except in the instance of differences between normal and abnormal groups. These differences are significant at the .05 level of probability. Variations not due to difference in psychiatric classification are due to individual differences.  $F$ 's are so small as to leave no doubt that the hypotheses must be accepted that differences due to color of ink used, order in which the sample is written, interaction between color and order, interaction between order and psychiatric classification, interaction between

ink color and psychiatric classification, and interaction between ink color, order of writing and psychiatric classification are those which might be expected by chance from a random sample of handwritings.

It would be interesting to discover what it is that contributes to the significant differences which exist between normal and psychiatric handwriting samples. However, the design of the present experiment does not permit inquiry into this matter.

### Summary

Color of ink employed in handwriting has no influence on the size of the handwriting.

Popular concepts concerning the influence of colored equipment or colored lights on motor performance (and possibly on emotional affect) must be revised until or unless more substantial evidence is uncovered to support these ideas. Nothing in the present experiment supports occupational therapy based on the influence of single colors.

Received December 13, 1952.

Early publication.

### References

1. Birren, E. *Color psychology and color therapy*. New York: McGraw-Hill, 1950.
2. Goldstein, K. Some experimental observations concerning the influence of colors on the function of the organism. *Occup. Ther. Rehabil.*, 1942, 21, 147-151.
3. Kravkov, S. V. Color vision and the autonomic nervous system. *J. opt. Soc. Amer.*, 1941, 31, 335-337.
4. Lukens, N. M., and Sherman, I. C. The effect of color on the output of work of psychotic patients in occupational therapy. *Occup. Ther. Rehabil.*, 1940, 20, 121.
5. Orr, M. E. Color therapy. *Occup. Ther. Rehabil.*, 1942, 21, 33-40.
6. Podolsky, E. *The doctor prescribes colors*. New York: National Library Press, 1938.
7. Prescott, B. D. The psychological analysis of light and color. *Occup. Ther. Rehabil.*, 1942, 21, 135-146.
8. Reeder, J. E. The psychogenic color field. *Amer. J. Ophthal.*, 1944, 27, 358-361.
9. Vollmer, H. Studies in the biologic effect of colored light. *Arch. Phys. Ther.*, 1938, 19, 197-211.
10. Mass. Assoc. for Occup. Ther. Bull., 1938, 12, 6-7.

## A Punched Card Procedure for Use with Partial Pairing

James E. Oliver

Cadillac Motor Car Division,  
General Motors Corporation,  
Detroit, Mich.

In using the method of paired comparisons, McCormick and his students (2, 3) have drawn attention to the feasibility of using partial pairing, as opposed to complete pairing, and have reported its use relative to the rating of employees. The partial pairing technique should, under any circumstance, result in the abbreviation of the time required for the preparation, rating, and scoring of pairs in proportion to the extent that pairing is partial rather than complete.

In a previous article (1) a procedure was discussed for the use of punched card equipment to facilitate rapid preparation and scoring of a complete pairing deck in accord with the traditional use of the paired comparison technique. This procedure can be tenably applied with equal facility to prepare a partial pairing deck. It should be particularly useful in cases where the method is used with  $N$ 's of 15-20 or greater.

### Partial Pairing

If  $N$  is an even number, the minimum number of pairs needed in a partial pairing deck is that required to give each of  $N$  individuals *opportunity* to receive at least one choice. The minimum number of pairs needed when  $N$  is an odd number, however, is that required to give each of  $N$  individuals *opportunity* for two choices. The composition of such a minimum partial pairing deck has been described by what Kephart and Oliver (1) have arbitrarily termed "set." Departure from complete pairing is conditioned by the number of "sets" incorporated in a partial pairing deck. The number of "patterns" (2) that may be used with any particular  $N$  is the number of possible combinations of "sets." In this respect, of course, the inclusion of all sets results in complete pairing.

If  $N$  is an even number, there are  $N/2$  sets in a complete pairing deck. As an example,

consider an  $N$  of 6, permitting numbers to represent names being paired.

Set 1	Set 2	Set 3
1 - 2	1 - 3	1 - 4
2 - 3	2 - 4	2 - 5
3 - 4	3 - 5	3 - 6
4 - 5	4 - 6	4 - 1
5 - 6	5 - 1	5 - 2
6 - 1	6 - 2	6 - 3

DESTROY

One-half of set 3 is destroyed since each half contains the same three pairs and is extraneous to a complete pairing deck. The three remaining pairs in set 3 give each of the six individuals opportunity for one choice. Set 1 and set 2, each composed of 6 pairs, give each of the 6 individuals opportunity for 2 choices. Therefore, we can pair everyone with one other individual by using only set 3, everyone with 2 other individuals by using either set 1 or 2, everyone with 3 other individuals by the combined use of set 3 with either 1 or 2, everyone with 4 other individuals by the combined use of set 1 and 2, or complete pairing by the use of all three sets. The small  $N$  of 6 is used for illustrative purposes only, but the same principle is operative for an even  $N$  of any size. For example, if  $N$  were 50, we would have 25 sets. Pairing can be made partial in multiples of 1, and the extent to which it is partial is determined only by the number of sets incorporated in the final deck to be used.

If  $N$  is an odd number, there are  $(N - 1)/2$  sets. As an example, consider an  $N$  of 7, again permitting numbers to represent individual names in the pairs.

Set 1	Set 2	Set 3
1 - 2	1 - 3	1 - 4
2 - 3	2 - 4	2 - 5
3 - 4	3 - 5	3 - 6
4 - 5	4 - 6	4 - 7
5 - 6	5 - 7	5 - 1
6 - 7	6 - 1	6 - 2
7 - 1	7 - 2	7 - 3

Although one-half of the last set is always destroyed when  $N$  is even, this is not characteristic when  $N$  is odd. Each of the three sets above consists of 7 pairs, and gives each individual opportunity to receive 2 choices. We can, therefore, incorporate either set 1, 2, or 3 into a partial pairing deck and have everyone paired with 2 other individuals, or use any two sets to pair everyone with 4 other individuals. The use of all 3 sets results in complete pairing. Therefore, when  $N$  is odd, pairing can be made partial in multiples of 2, and the extent to which it is partial is determined by the number of sets incorporated in the deck.

### Summary

The method of paired comparisons has long been considered somewhat laborious to say the least. In a previous article (1) a punched card procedure was outlined to facilitate rapid preparation and scoring of the

pairs as the method has been traditionally used. The discussion of the punched card procedure has here been extended to draw attention to its applicability to partial pairing, a technique to further abbreviate time and labor requirements in preparing, rating, and scoring the pairs. The procedure is systematic and may be used with any number of variables.

*Received June 7, 1952.*

### References

1. Kephart, N. C., and Oliver, J. E. A punched card procedure for use with the method of paired comparison. *J. appl. Psychol.*, 1952, 36, 47-48.
2. McCormick, E. J., and Bachus, J. A. Paired comparison ratings. I. The effect on ratings of reduction in the number of pairs. *J. appl. Psychol.*, 1952, 36, 123-127.
3. McCormick, E. J., and Roberts, W. K. Paired comparison ratings. II. The reliability of ratings based on partial pairings. *J. appl. Psychol.*, 1952, 36, 188-192.

## Pointer Location and Accuracy of Dial Reading

Sherman Ross,

William Ray

and

Louis Della Valle

*University of Maryland*

Accuracy of dial reading and the conditions of which it is a function constitute a problem of interest to those psychologists who are concerned with display problems. The facts of the relationship between accuracy and its determinants have important applications to industrial and military situations. Reviews of the previous work accomplished have been presented in several sources (1, 3, 5, 8).

Although a considerable amount of experimental effort has been expended in this area, little attention has yet been paid to the specific questions with which the present study was concerned. In general, this experiment attempted to determine the relationship between the accuracy of reading and the dial sector and specific location of the dial pointer.

Kappauf and Smith (7) found that the sector had no consistent effect on either local errors or systematic errors for many dials, but sector location may influence the occurrence of specific systematic errors on certain scales. Dials graduated from 0 to 50 and 0 to 100 revealed an error more prevalent on right dial halves than on left halves on scales numbered by tens.

Christensen (2) studied exposure time as a factor in dial reading performance. Moving scale dials were better at short exposures while moving pointer dials were better at long exposures. Sleight (9) compared dial shapes for legibility. In the order of accuracy of readings the dials ranked as follows: (1) open-window; (2) round; (3) semicircular; (4) horizontal; and (5) vertical.

In a study of instrument recording performance under varied illuminating conditions, Spencer (10) reported readings most accurate at the 12 o'clock sector of the dial, but his results were not consistent. In stud-

ies of check reading of fixed-scale, moving pointer instruments, Warrick and Grether (11) and Grether and Connell (4) reported more frequent correct responses when the index is at the 9 o'clock position than when it is at the 3 o'clock position.

In a study of the effect of pointer design and pointer alignment position on speed and accuracy of instrument readings, White (12) had his subjects make a qualitative reading of the deviation from vertical among 16 simulated engine instruments in order to make a correction. Alignment at the 9 o'clock position was superior for qualitative reading. In another experiment his subjects had to check-read a panel of simulated instruments with pointer alignment at the 9, 12, 3, and 6 o'clock positions and indicate misalignment. No significant differences in response time and errors were found. Horton (6) found an increase in the frequency of systematic errors with sector errors being more than twice as frequent on the left half of the scope as on the right. In an unpublished study from this laboratory it was found that fewer errors were made at and around the 9, 12, and 3 o'clock positions and more errors were made at some intermediate points in a circular dial. Our results in this respect were not entirely consistent, however, because in both groups there were mid-division settings which were not numbered.

From the literature cited several findings are of particular interest in connection with the present study. Kappauf and Smith (7) found that sector had no consistent effect. When reversal errors were frequent, sector was then observed to be important. Spencer (10) reported more accurate readings in the 12 o'clock sector, but his results were not consistent. White (12) found that the 9

o'clock position was superior for reading of deviations from vertical. Finally, we observed a tendency for fewer errors at and around the 9, 12, and 3 o'clock positions.

Three dial shapes were used in the present study in an attempt to answer certain questions which can be raised concerning the influence of sector and pointer location on accuracy of reading. These dials are: (A) *semicircular upright dial*; (B) *semicircular inverted dial*; and (C) *circular dial* (see Figure 1).

The following specific questions were asked concerning accuracy of reading the three dials:

1) Are errors in a particular quadrant a function of the dial shape in which the quadrant occurs?

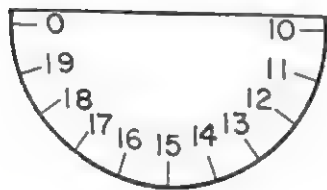
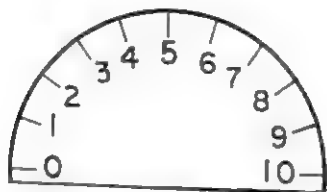


FIG. 1. The dial shapes used in the experiment.

2) Are intra-dial errors for Dial C a function of the quadrant in which readings are made?

3) Are intra-dial errors for Dial C related in a systematic way to pointer positions of 9, 12, 3, and 6 o'clock compared to intermediate positions?

4) Are errors a function of the dial half (upper and lower) in which readings are made?

### Method and Procedure

**Subjects:** The subjects used in the experiment were eight male and two female university students. They ranged in age from 20 to 30 years. Each subject had a minimum Snellen index of 20/20 (corrected or uncorrected) in each eye.

**Apparatus:** The apparatus used to present the dial settings to the subjects was a modification of the Dodge tachistoscope. The interior was painted black, and the subject viewed a single dial through a binocular eyepiece. The pre-adapting illumination and the presentation illumination were provided by two pairs of 25 watt bulbs. The distance from the subject's eyes to the test dials was 42 in.

An electronic interval timer was used to present exposure periods which were set at 0.1, 0.3, 0.4, 0.5, and 0.7 sec.

The dials were constructed to follow the design characteristics suggested in "Standards to be Employed in Research on Visual Displays," Armed Forces-NRC, Vision Committee, 1 March 1950. All characteristics of the three dials were held constant as shown below:

1. All numbers were made by India Ink on white cardboard using a No. 3 pen and a LeRoy lettering guide.

2. The diameter of each dial was  $2\frac{1}{2}$  in.

3. The distance between graduations along the circumference of the scale was  $\frac{3}{8}$  in. and the length of each graduation unit was  $\frac{3}{16}$  in.

4. The height of each numeral was  $\frac{5}{32}$  in., and the stroke width was approximately  $\frac{1}{32}$  in.

5. The 0 setting for each dial was at 9 o'clock and the 10 setting was at 3 o'clock.

Table 1

The Error Scores for the Three Dials Tested

Subject	Dial A			Dial B			Upper Half	Lower Half	Dial C				Cardinal Points	Intermediate Points
	Whole	Q I	Q II	Whole	Q III	Q IV			Q I	Q II	Q III	Q IV		
1	35	13	22	11	3	8	18	27	7	11	14	13	0	6
2	5	3	3	9	3	6	15	21	7	8	14	7	0	2
3	17	7	10	15	7	9	17	18	8	11	7	11	2	10
4	15	5	10	11	5	6	19	10	6	13	7	3	0	4
5	14	4	10	3	2	1	16	19	6	10	11	8	0	8
6	12	3	9	24	11	13	22	25	9	13	17	8	1	11
7	13	3	10	18	8	10	32	24	16	16	16	7	1	13
8	26	15	11	11	4	7	15	21	10	5	10	11	0	11
9	4	1	3	15	10	6	7	9	4	3	1	9	1	6
10	35	14	21	34	15	19	17	16	7	10	9	9	2	6

6. The pointer was  $1\frac{3}{16}$  in. long and was  $\frac{1}{8}$  in. wide.

Each dial was mounted on stiff black cardboard. The settings on the test dial were manipulated by means of a larger dial placed on the reverse side of the test dial. Thus, settings on each dial could be quickly and conveniently changed.

*Procedure:* After the subject was seated before the binocular eyepiece of the tachistoscope, a dial was exposed for an unlimited exposure. The subject was shown the dial, and its units and graduations were pointed out. The subject was shown several settings, and was told that he would be required to report the pointer position. The pointer was set either on the graduation marker or midway between two graduation markers. The subject was also shown the dial under the conditions of timed exposures. The experimenter called "ready" when a trial was to be started, and the click of the interval timer signaled the starting of the timed exposure. The subject reported "11," "9 $\frac{1}{2}$ ," "6 $\frac{1}{2}$ ," "17," etc. Dials A and B had 21 possible settings while Dial C had 40 possible settings.

The order of presentation of dials, the setting on each dial, and the time interval were systematically varied in order to handle possible practice and fatigue effects. Dial A and Dial B were each presented 105 times for each subject involving the 21 different dial settings and the five time intervals tested.

Dial C was presented a total of 200 times to each subject. Thus each subject made 410 judgments involving the three dial faces tested, and the results presented are based on a total of 4,100 judgments.

### Results and Discussion

The error score for a given individual for any set of dial readings was found by summing twice the deviation from the actual setting. Thus, score =  $\Sigma (2E)$ , where E is the deviation of the subjects reading from the actual dial setting. Each deviation was multiplied by two simply to eliminate decimals. These results are shown in Table 1 for each of the three dials tested. The table shows the total error score for each individual for comparable sections, dial-wise or quadrant-wise, for Dials A, B, and C. In addition, for Dial C the total error score is shown for cardinal settings (0, 5, 10, and 15) and for intermediate settings (2, 3, 7, 8, 12, 13, 17, and 18). The quadrants referred to are designated as follows: (I) upper-right, (II) upper-left, (III) lower-left, and (IV) lower-right.

For any statistical test of significance a difference score was found for each individual. The standard error was then computed from the distribution of differences, thus allowing for the correlation among individuals. The results of the tests of significance (*t* test) are shown in Tables 2; 3, and 4.

Table 2

Significance of Difference Between Error Scores in Comparable Quadrants of Dials A, B, and C

Quadrant	Dial Comparison	<i>t</i> -value
I	A vs. C	0.62
II	A vs. C	0.45
III	B vs. C	1.73
IV	B vs. C	0.01

Table 3

Significance of Differences Between Error Scores Made in Quadrants of Dial C Expressed as *t*-values

Quadrant	I	II	III	IV
I	—			
II	1.89	—		
III	2.13	0.48	—	
IV	0.44	0.81	1.14	—

Table 4

Significance of Differences in Total Error Scores in Dials A, B, and C Expressed as *t*-values

Dials	A	B	C (upper half)
A	—		
B	0.69	—	
C (upper half)	0.05	0.90	—
C (lower half)	0.41	1.16	0.65

Four sets of *t* tests were made relative to the four questions previously raised. The questions are restated here, as follows:

1) Are errors in a given quadrant a function of the dial shape in which the quadrant occurs?

2) Are intra-dial errors (Dial C) a function of the quadrant in which readings are made?

3) Are intra-dial errors (Dial C) related in a systematic way to pointer positions of 0, 90, 180, and 270 degrees compared to intermediate positions?

4) Are errors a function of the dial half (upper vs. lower) in which readings are made?

From the results of the tests of significance shown in Table 2, it is clear that dial shape

for the three dials used in the experiment has not been demonstrated to be an important factor in reading accuracy when the errors produced are considered on a quadrant basis, since all of the *t*-values shown are insignificant at the 5 per cent level of confidence.

From the results shown in Table 3 dealing only with the errors produced in the circular dial (Dial C), it may be concluded that the quadrant from which the settings are read has not been demonstrated to be a significant factor in reading accuracy, since all of the *t*-values shown are insignificant at the 5 per cent level of confidence.

The third major comparison to be considered in the analysis of the data is the result of the comparison of error performance when errors made at dial settings 0, 5, 10, and 15 are compared with errors made at settings 2, 3, 7, 8, 12, 13, 17, and 18 for the circular dial (Dial C). The *t*-value here is 5.89 and the difference is significant at the .01 level. We, therefore, conclude that in the circular dial used in the study significantly fewer errors were made at the 9, 12, 3, and 6 o'clock positions than at the tested intermediate points.

Table 4 shows the comparison of the accuracy of reading the upper half of Dial C with Dial A, the lower half of Dial C with Dial B, Dial A with Dial B, etc. Here again none of the *t*-values are significant at the 5 per cent level of confidence. The results show that accuracy of reading in upper and lower dial halves does not differ significantly in the set of dials used in this study.

One additional finding should be noted. The results of previous investigations (2) concerning the effect of exposure time were verified. Errors decreased as length of exposure time increased.

### Summary

The purpose of this experiment was the determination of the relationship between accuracy of dial reading and the sector and specific location of the dial pointer. The three dials used were a semicircular upright dial, a semicircular inverted dial, and a circular dial. Ten subjects made a total of

4,100 judgments at five exposure times on the three dials.

Tests of significance for error scores were made and permitted the following conclusions:

1) Differences in dial shape were not an important source of error.

2) Differences in sector location of the dial pointer were not an important source of error.

3) Significant differences in error scores were found for readings made at 9, 12, 3, and 6 o'clock positions corresponding to pointer settings at 0, 5, 10, and 20 when compared with intermediate points.

4) No significant differences in error scores were found when upper and lower dial halves were compared.

These findings suggest that critical regions of a scale should be assigned to the 9, 12, 3, or 6 o'clock positions of a circular dial, and that factors other than errors may be considered in the choice of a dial from among the three types studied here.

Received May 22, 1952.

# References

1. Chapanis, A., Garner, W. R., and Morgan, C. T. *Applied experimental psychology*. New York: Wiley, 1949.
2. Christensen, J. M. Exposure time as a factor in dial reading performance. *Amer. Psychologist*, 1951, 6, 387.

3. Fitts, P. M. Engineering psychology and equipment design. In *Handbook of experimental psychology*. S. S. Stevens (Ed.). New York: Wiley, 1951.
4. Grether, W. F., and Connell, S. C. Psychological factors in check reading single instruments. *USAF Memo. Rept. No. MCREXD-694-17A*, USAF, Air Materiel Command, 20 September 1948. Pp. 21.
5. *Handbook of human engineering data for design engineers*. Medford, Mass.: Tufts College, 1949. Chap. IV, Sec. I.
6. Horton, G. P. An analysis of errors made in a schematic PPI display. *USAF Tech. Rept. No. 5960*, USAF, Air Materiel Command, October, 1949.
7. Kappauf, W. E., and Smith, W. A. Design of instrument dials for maximum legibility III. Some data on the difficulty of quantitative reading in different parts of a dial. *USAF Tech. Rept. No. 5914*, Part 3, May, 1950.
8. McFarland, R. A. *Human factors in air transport design*. New York: McGraw-Hill, 1946.
9. Sleight, R. B. The effect of instrument dial shape on legibility. *J. appl. Psychol.*, 1948, 32, 170-178.
10. Spencer, J. Presentation of information by aircraft instruments. II. Instrument recording performance under varied illuminating conditions. *Flying Personnel Res. Comm.*, No. 754, May, 1951.
11. Warrick, M. J., and Grether, W. F. The effect of pointer alignment on check reading of engine instrument panels. *USAF Memo. Rept. No. WCREXD-694-17*, USAF, Air Materiel Command, 4 June 1948.
12. White, W. J. The effect of pointer design and pointer alignment position on the speed and accuracy of reading groups of simulated engine instruments. *USAF Tech. Rept. No. 6014*, USAF, Air Materiel Command, July, 1951.

## Dimensional Analysis of Motion: V. An Analytic Test of Psychomotor Ability<sup>1</sup>

Shelby Harris and Karl U. Smith

*University of Wisconsin*

The present paper describes a new test of psychomotor skills, based on dimensional and component analysis of movements in motion. This test, which has been named the Analytic Reactometer, permits separate and automatic registration of the travel and manipulation components of motion involved in the successive grasping and manipulating of objects.<sup>2</sup> This development in psychomotor testing is of considerable significance for several reasons: (1) the test provides more detailed and precise measures of the components of motion than has been previously possible; (2) it permits, within the same instrument, systematic variation of several dimensions of motion, such as extent of movement, direction of movement, extent of manipulation, complexity of movement and manipulation, plane of movement, hand involved, etc.; (3) it provides a means of analyzing errors of manipulation in terms of the various dimensions of motion and with regard to the component time scores; and (4) the principles involved in the test may be incorporated in all types of psychomotor tests which may be designed to simulate various types of work situations. The desirability of a psychomotor test situation which will accomplish the objectives named is indicated by the fact that different components and dimensions of movement in skilled motion patterns are functionally distinct (1, 2, 3, 5, 6) and often uncorrelated. The test to be described has some significance for the field of personnel selection, but it is believed that the methods and results described herein are more immediately applicable to problems of detailed measurement of different types of

human manual performance in medical and industrial research.

### Methods

The Analytic Reactometer is designed in terms of two main features: (1) control of the space dimensions of the motion pattern; and (2) separate measurement of the manipulative and travel components of motion.

The planned performance situation used in the present form of the test is a control panel 45.7 cm. square, on which are mounted 25 rotary switches in 5 rows of 5 switches, each spaced 7.6 cm. apart (Figure 1). Each switch has 17 settings, selected points of which are marked as shown in Figure 1. The positions thus marked are 40°, 80°, and 180° clockwise and 40° and 80° counter-clockwise.

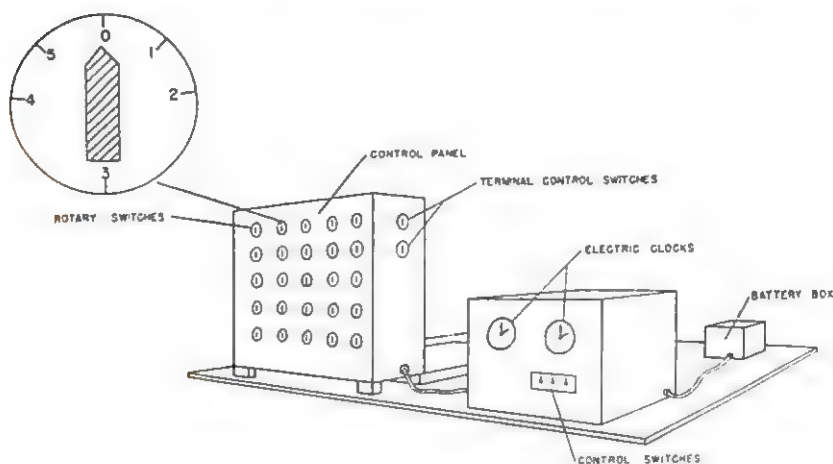
The manipulative and travel components of motion are measured separately in this test by means of an electronic motion analyzer (4), consisting of a balanced relay circuit, in which the subject acts as a key. When the subject touches one of the switches, the analyzer is activated and elapsed time is recorded on a precision time clock<sup>3</sup> in hundredths of a second until contact with this switch is broken. When the clock measuring manipulation time stops, a second clock, measuring travel time, starts, and continues to run until the next switch is touched. Thus, the elapsed time in operating any pattern of the switches is totalled separately for manipulation and travel movement components by means of the two clocks.

The following types of scores may be obtained on the test: (1) time involved in turning the 25 switches; (2) time of travel between the 25 switches; (3) total time involved in both manipulation and travel; and

<sup>1</sup> This study has been supported by funds voted by the Legislature, The State of Wisconsin, and assigned by the Graduate School Research Committee, The University of Wisconsin.

<sup>2</sup> The analysis of results of this study has been aided by the facilities of the Computing Service, The University of Wisconsin.

<sup>3</sup> Model S1, Standard Time Clock, Standard Electric Time Company, Springfield, Massachusetts.



A SCHEMATIC DIAGRAM OF THE ANALYTIC REACTOMETER

FIG. 1. Diagram of Analytic Reactometer showing the arrangement of controls on the panel and the timing mechanism. The inset illustrates the design of each manual control. The special mount for the control panel makes it possible to position the panel in different planes.

(4) errors made in positioning the switches. The reactometer permits testing of the performance of either hand, with different planes, directions, and magnitudes of movement. To vary these dimensions of motion, different settings and patterns of switches may be used or the control panel itself may be changed from one plane to another. The whole test is constructed to be easily transported.

The main objective of this study has been to analyze, in terms of correlation procedures, the interrelations between different reactive

variables which typically enter into performance on psychomotor tests. Specifically, the reliability of scores related to different dimensions and components of motion, as performed in the test situation, has been determined. In addition, intercorrelations between the components of motion and between tests involving different dimensions of motion have been computed.

Twenty tests were carried out in the study which covered the following aspects of motion: (1) right and left directions of travel movement; (2) different directions of movement; (3) performance with each hand; (4) horizontal and vertical planes of motion; and (5) simple and complex patterns of manipulation. All switches on the board

were used in each test. In all of the tests the manipulative movement consisted of a  $40^\circ$  rotation of the switch, either right or left to positions 1 or 5 respectively. The travel movement from switch to switch was horizontal (left to right) in some tests and vertical (downward) in other tests. Complex manipulation patterns differed from the simple ones in that alternate switches were turned in opposite directions. It was not feasible to use a balanced sequence of the different tests to control practice effects. Instead, for this preliminary study, the twenty different tests were administered to all subjects in the same sequence.

A total of 78 college students served as subjects in the study. All 20 tests described above were given to each subject. Each test required approximately one minute to administer. The subject was instructed to return the switches on the panel in the pre-defined patterns as rapidly as possible and at the same time to be careful to position each switch accurately. When a new general pattern of motion was introduced, the subject was given a practice trial on the first ten switches to be turned in this pattern of motion. Reliability figures and intercorrelations between components and dimensions of motion were computed not only for each in-

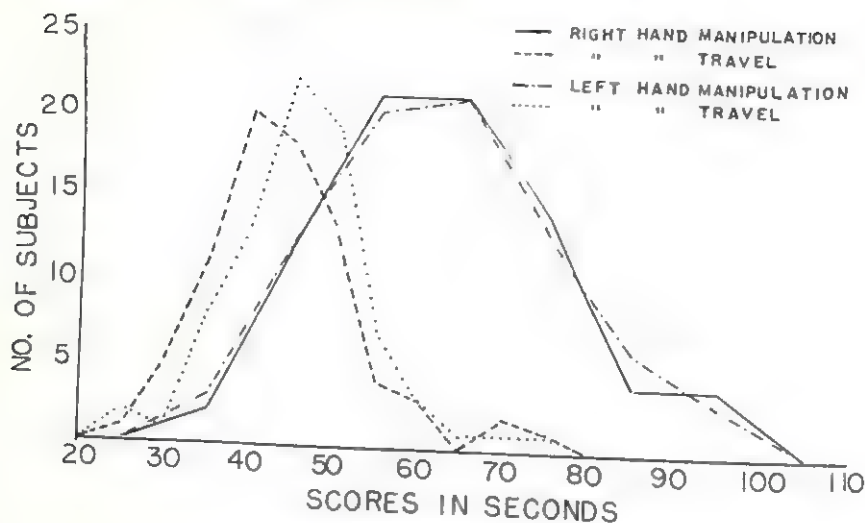


FIG. 2. Distributions of manipulation and travel scores in the simple manipulation pattern for the right and for the left hand in 78 subjects. The distribution of scores for manipulation by the two hands are identical, whereas the travel scores show some discrepancy between the hands. Both pairs of distributions are slightly skewed positively.

dividual test but also for combined scores of the various tests involving a common dimension of motion. Of the 78 subjects, 49 repeated the sequence of tests some 10 to 14 days after the initial administration. Data obtained on these subjects are used to compute the test-retest reliability of the different measures obtained on the Reactometer.

### Results

Typical distributions of test scores are presented in Figures 2 and 3. The distributions for the right and left hands shown in Figure 2 are based on combined scores for all tests involving simple manipulation patterns. There were 8 of these tests for each of the hands. Figure 3 shows the analogous

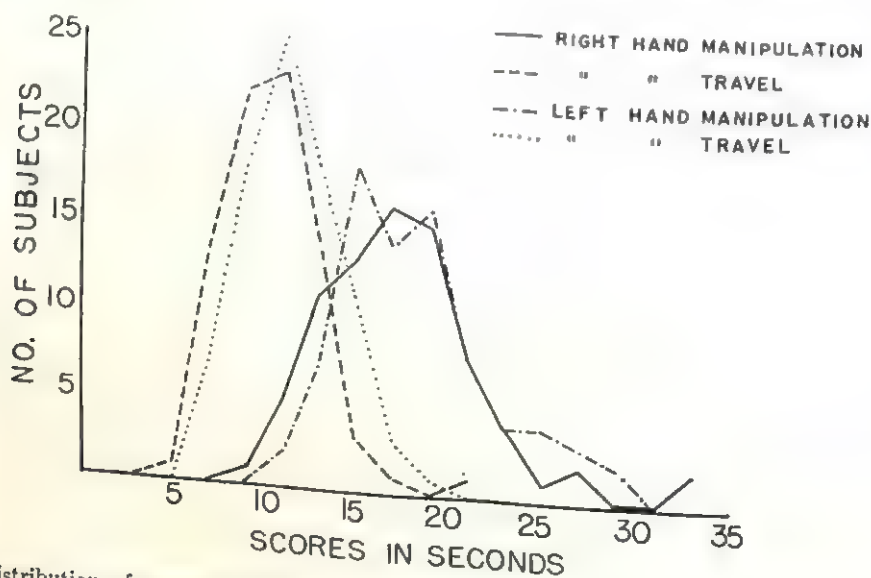


FIG. 3. Distribution of scores for 78 subjects in the complex manipulation patterns for travel and manipulation in relation to the two hands. The pattern of scores found for the simple manipulation patterns appears also in these complex patterns.

Table 1

Means and Standard Deviations for Different Combined Tests on the Analytic Reactometer

Test	First Test				Second Test			
	Manipulation		Travel		Manipulation		Travel	
	M	$\sigma$	M	$\sigma$	M	$\sigma$	M	$\sigma$
Hor. Plane	68.5	17.4	42.2	9.2	58.1	13.1	41.2	7.2
Ver. Plane	59.1	14.2	40.2	8.5	53.0	13.0	41.1	7.9
Right Hand	63.4	15.0	40.2	8.9	54.8	12.7	40.1	7.9
Left Hand	64.2	16.0	42.2	8.6	56.3	13.2	42.2	7.5
Lat. Direction	65.3	15.7	43.9	9.4	55.4	12.5	43.6	8.1
Ver. Direction	62.3	15.3	38.5	8.1	55.7	13.1	38.7	7.2
Manip. Right	63.9	15.4	41.0	8.8	55.6	12.8	41.3	7.8
Manip. Left	63.8	15.3	41.3	8.4	55.5	12.8	41.0	7.5
Total Simple	127.7	30.5	82.3	17.1	111.1	25.5	82.3	15.1
Total Complex	36.1	8.6	20.5	4.2	32.0	7.8	20.5	3.6

Table 2

Test-Retest Reliability of the Component Tests with Respect to Both Manipulation and Travel Scores

Note: Each test was of one minute duration.

	Manipulation	Travel
I. Simple Manipulation		
A. Horizontal Plane		
1. R. H., R. Manip., Trav. R.	.64	.60
2. R. H., L. Manip., Trav. R.	.53	.50
3. L. H., R. Manip., Trav. R.	.71	.76
4. L. H., L. Manip., Trav. R.	.83	.58
5. R. H., R. Manip., Trav. In	.68	.81
6. R. H., L. Manip., Trav. In	.66	.50
7. L. H., R. Manip., Trav. In	.78	.70
8. L. H., L. Manip., Trav. In	.73	.57
B. Vertical Plane		
9. R. H., R. Manip., Trav. Right	.74	.71
10. R. H., L. Manip., Trav. Right	.70	.65
11. L. H., R. Manip., Trav. Right	.81	.72
12. L. H., L. Manip., Trav. Right	.70	.59
13. R. H., R. Manip., Trav. Down	.77	.58
14. R. H., L. Manip., Trav. Down	.78	.67
15. L. H., R. Manip., Trav. Down	.79	.69
16. L. H., L. Manip., Trav. Down	.74	.66
II. Complex Manipulation		
A. Vertical Plane		
17. R. H., R-L Manip., Trav. Right	.74	.68
18. L. H., R-L Manip., Trav. Right	.77	.66
19. R. H., R-L Manip., Trav. Down	.78	.44
20. L. H., R-L Manip., Trav. Down	.76	.61

Table 3

Test-Retest Reliability for Various Combined Scores

	Manipulation	Travel
Right Hand	.81	.75
Left Hand	.87	.77
Right Manipulation	.85	.81
Left Manipulation	.84	.69
Lateral Travel	.83	.76
Down and In Travel	.85	.75
Horizontal Plane	.81	.73
Vertical Plane	.86	.75
Total Simple Manipulation	.86	.77
Total Complex Manipulation	.86	.69

distributions for combined scores of four tests involving complex manipulation patterns. It may be seen from these distributions that both the manipulation time and travel time distributions are similar for the two hands. All of the distributions approach normality.

The test and retest means and standard deviations for various combined scores are given in Table 1. Each of the combined scores is based on the performance of 49 subjects for all of the tests which involved the specified dimension. With the exception of the compound score of all tests involving complex manipulation, all of these figures

are based on the tests involving simple manipulation patterns. Comparison of the means for various dimensions of motion is not justified due to the lack of control over practice effects.

Tables 2 and 3 present the test-retest reliability figures for the twenty individual tests and for the various combined scores. The reliability figures for the individual tests are presented in the order that the tests were administered. The combined scores on which the reliability figures in Table 3 are based are the same as those of Table 1. All of the reliability values are relatively high. The manipulation-time coefficients are consistently higher than the travel-time values.

Table 4 shows the correlations between the manipulation and travel components of motion for the combined scores and the correlations between the several dimensions of motion involved in the study. All of these figures, which are based on data of 78 subjects, are positive coefficients. It is obvious from the table that the relationship between the components of motion is consistently low. Nine of these coefficients are significantly different from zero at the one per cent level. One is significant at the five per cent level. The correlations between dimensions of motion are high for both manipulation and

Table 4

Correlations Between Components and Dimensions of Motion

	Man. vs. Trav.	Correlation between Dimensions	
		Manipulation	Travel
Right Hand	.37**		
Left Hand	.31**	.90**	.81**
Right Manipulation	.34**		
Left Manipulation	.35**	.98**	.94**
Lateral Travel	.41**		
Down and In Travel	.29**	.92**	.96**
Horizontal Plane	.39**		
Vertical Plane	.29**	.85**	.88**
Total Simple Manipulation	.36**		
Total Complex Manipulation	.25*	.93**	.85**

\* Significant at 5% level.

\*\* Significant at 1% level.

travel components of motion. Among the correlations between dimensions, the values for planes of motion are somewhat lower than those for other dimensions. Generally, the intercorrelations between dimensions of motion are higher for the manipulative aspects of motion than those for the travel components.

### Summary

A special psychomotor test for separate measurement of the travel and manipulation components of motion has been described. The test, called the Analytic Reactometer, permits controlled variation and measurement of different bodily and space dimensions of motion which are involved in various types of motion patterns.

Preliminary investigation employing the instrument have yielded the following general results:

1. Critical sources of variation in performance in various motion patterns of the type studied are related to the manipulation and travel components of motion.

2. Performances in different space dimensions of both manipulation and travel movements correlate highly with one another.

3. The reliability of specific tests related to hands, planes, direction of travel, direction of manipulation and complexity of the manipulation pattern in the general test situa-

tion described typically exceeds  $+ .80$  for manipulation and  $+ .75$  for travel movements.

4. The present test, and the principles behind it, provide one means of securing precise and analytical data for exact quantitative specification of motions and motion functions. Application of analytical methods described to studies of growth, aging, neurological deficiency, and to industrial selection may advance considerably the scientific validity of data concerning human motion.

Received June 13, 1952.

### References

1. Davis, R., Wehrkamp, R., and Smith, K. U. Dimensional analysis of motion: I. Effects of laterality and movement direction. *J. appl. Psychol.*, 1951, 35, 363-366.
2. Lincoln, R. S., and Smith, K. U. Systematic analysis of factors determining accuracy in visual tracking. *Science*. In press.
3. Rubin, G., Von Trebra, P. A., and Smith, K. U. Dimensional analysis of motion: III. Complexity of movement pattern. *J. appl. Psychol.*, 1952, 36, 272-276.
4. Smith, K. U., and Wehrkamp, R. A. Universal motion analyzer applied to psychomotor performance. *Science*, 1951, 113, 242-244.
5. Wehrkamp, R., and Smith, K. U. Dimensional analysis of motion: II. Effects of travel distance. *J. appl. Psychol.*, 1952, 36, 201-206.
6. Von Trebra, P. A., and Smith, K. U. Dimensional analysis of motion: IV. Transfer effects and direction of movement. *J. appl. Psychol.*, 1952, 36, 348-353.

## Applied Psychology in Action

*Editor's Note:* With this issue, we begin what may become a regular feature of the *Journal of Applied Psychology*. We plan to publish brief descriptions of *applied psychology in action* to be written by psychologists who are applying psychology in real life situations. Brief news notes concerning applied psychology in action from a variety of sources will be published. Descriptions of procedures and techniques believed to be effective, even though desirable *experimental*

*controls* may not have been possible, will be included. Thus, a forum for the interchange of practical information will be provided practitioners of applied psychology. In part, this new feature of the *Journal of Applied Psychology* attempts to meet the challenge contained in Dr. Marion A. Bills' presidential address before the Division of Industrial and Business Psychology last September. It is appropriate, therefore, to begin with the publication of her provocative address.

THE JOURNAL OF APPLIED PSYCHOLOGY  
Vol. 37, No. 2, 1953

### Our Expanding Responsibilities \*

Marion A. Bills

*Aetna Life Insurance Company, Hartford, Connecticut*

Three items lead me to the choice of the title for this talk: (1) the most interesting diaries which many of the psychologists who are working full time in industry have kept for two weeks and sent in as a foundation for a case book in industrial psychology; (2) a meeting which I attended of psychiatrists and psychologists working in industry which was held in Asbury Park this spring; and, (3) the criticisms which have been made, some in writing and many in discussions, that our published research is at a very superficial level.

The diaries which we have received from individuals in private industry indicate clearly that our duties spread over the entire field of management. There was about an even division between duties which are involved in the setting of policies and those which have to do with the administration of those policies.

Some of the diaries indicated concentration of effort in a given field. Almost the full two weeks of one psychologist's time was spent on where to locate a new plant with emphasis on labor procurement. The diary

ended, "If I sent you one three months from now it would be entirely different." With Wage Stabilization still in force how to keep within the law and still run a business is occupying 70 per cent of one psychologist's time in a mid-western company. One person delayed sending in his diary until union negotiations were over because he had done nothing "psychological" (this is a direct quotation) for the month he had been handling the negotiations.

Many of the diaries varied from day to day. They included conferences (we seem to run to conferences) on wage systems, including merit rating, conferences on training ranging from salesmen to supervisors to hourly workers and including such detailed items as the purchase of an opaque projector for use in safety training—discussions on the editorial policy of a house organ—conferences on pension systems, and how to prepare the individual for retirement and some actual work with the individuals—attendance at a meeting on a proposed Stock Purchase Plan and so through the entire range of management. Throughout most of the diaries was an occasional hour or two spent directing or actually doing work on research prob-

\* Presidential address delivered before the Division of Industrial and Business Psychology at the 1952 APA meeting in Washington, D. C.

lems and one sensed that in many instances there was a desire to do more—time only being lacking. One of the diaries ended with an hour devoted to consulting with one of his assistants on a research problem of selection and then honesty prevailing he added a note, "This hour is really wishful thinking; it was only 15 minutes."

The great volume of managerial work that we do and which was clearly brought out in the diaries was pointed up for me at the meeting of the psychiatrists and psychologists at Asbury Park. You can count on the fingers of one hand the number of psychiatrists in private industry and of these few, only one was talking of management problems and he apologized for his interest. The psychiatrists, whether they be dealing with charwomen or the president of a company, were all talking of individuals as individuals. Our interest in groups and in organizations was entirely lacking among them. This is a ball which we are apparently carrying alone.

How and why have we gotten ourselves into this situation for we are in it much more than doctors, lawyers or engineers. First, I believe because we are a newer science and our field is much less defined. A problem must have at least a medical tinge before management goes to a medical department but psychology being a bit vague in the mind of management they feel free to turn to a psychologist on almost any problem. Second, because by and large we have felt rather complimented to get into many managerial functions and have taken them on willingly. Third, I believe the most important is that as we go into managerial work we carry with us many fundamental psychological principles, and so influence management in the way that as psychologists we feel they should be influenced and our influence is greater because we do not wear a tag which says "psychologists." What are these principles that we carry over? I believe one of the most important is the principle of "Stop, look and listen" that as scientists has been ground into us in all of our training. Management has long been accustomed to getting the facts on financial problems, on machine operation, on costs in factory upkeep, etc., but many of

their judgments on people have been on a random basis of single cases—rumor and prejudices. I remember 30 years ago one man who had built up a big business selling office machines told me that all black-haired men were dishonest; at that point no amount of pointing out honest black-haired men had any effect on his prejudice and yet almost any one of us given a year or two and some tact could have worn him down and at least improved his evaluation of personnel. In any company it is a long selling job that persons' reactions can be studied on a scientific basis—that persons can be selected for any job with a fairly accurate prediction of success or failure—that they will react in certain ways to certain types of training—that what they want can be determined—that fair wage rates can be established that will take into consideration the difficulty of the job and the efficiency of the individual and will cause at least 50 per cent of the personnel to say the company is fair. One has only to make one or two sales of this type and they may at the beginning take a long time until one becomes a part of management. This is what I think has happened to us. We have made the sales. As management has grown to realize that their personnel is their chief asset, the person that can tell them about that personnel has been drawn into decision making functions.

It's a long selling job because one must not only convince top management who probably were already favorable before we were hired, but we must sell the idea all the way down the line that the scientific approach is going to make each person's work more effective and take not a whit away from his own responsibilities.

We have learned a great deal over the years; perhaps more than we have given and much more than we realize, but the final result has been beneficial to both management and ourselves. Let us give an example. Our first study of the interview was a debunking of it. We showed very successfully that the average interview was a very weak tool for selection of personnel. For example—you all remember those experiments by *Hollingworth*, where 10 sales managers interviewed 20 men

and if each picked the two that he considered best, 18 would have been chosen. But where did this most interesting scientific experiment get us? Practically nowhere! As psychologists we got excited but sales managers said, "How interesting" and went right on picking salesmen by interviews only. Then, gradually we modified the approach. In substance we said, "The interview is the tool by which the final decision must be made but what information concerning the individual can we give you—the sales manager—that will help in this final decision?" With research we showed that some tests were helpful—that there were certain ways of scoring an application blank that came out with an indication of success or lack of success. Management bought the results and we had learned to play on the team and playing on the team we could gradually make suggestions which changed somewhat the type of interview and helped to make it as an interview more successful.

We have a mighty heritage of at least seventy-five years of psychological research back of us to which is constantly being added new and valuable data and ideas. Much of it is written down in our literature. Some has been handed to us by word of mouth from our professors and colleagues. It is well worth using and we are using it but the criticism that our publications as psychologists in private industry have been too few and too superficial is probably just.

For example; for at least three years many of us have been worrying about the frustrated foreman. I think we have done something about it in our own companies both on the policy setting level and with the individual foreman or supervisor, but it's the academic man who writes about it. We do not particularly like what he writes—he puts the blame too much on the foreman and seeing the many complications that the foreman is meeting, and because we like him as a friend, we become a bit resentful. We talk about the Ivory Tower, and we may even quote Kipling about "the butterfly along the road preaching contentment to the toad." But, we do not write about the frustrated foreman as we see him. We shake our heads, and

say the subject is too big or too hot, and what we write, if we write at all, is a small statistical study of how to select or rate the bench worker or the file clerk. Of course, being a psychologist I now modify my statement and say that there are exceptions, and some of these exceptions are outstanding. However, on my desk as I wrote this was the December, 1951 *Journal of Applied Psychology*, and the Winter Number of *Personnel Psychology* and with the everlasting compulsion of a psychologist to count (we all seem to have this compulsion), I counted the articles in these two magazines. A total of 14 articles were from persons connected with colleges—5 from the military force—2 from consulting firms—and one made up of two junior authorships from persons in private industry. I think this is fairly typical and certainly our showing is not good, and our friends in the consulting field although they did twice as well as we did, cannot pat themselves on the back too much either. Together we contributed only a sixth of the articles in the two magazines where you would expect us to make the best showing.

Why then the big difference between what we are doing and what we are reporting?

There are many reasons but may I illustrate a few and I am asking you to bear with me while I quote an experience of my own so far back that it no longer has a personal connotation. Twenty-five years ago we had an experiment on sound-proofing, by installing a sound-proof ceiling in a department for which we had good production records, and could continue those records after the installation. Based on the results we spent a half million dollars sound-proofing a new building. I am fully convinced that the decision was correct but I never published the results. The most amateur statistician could have shot them full of holes; a fear complex partly, but also a recognition that some results cannot be accurately measured. As one talks to psychologists in private industry one hears this often. I know one psychologist who has set up a new training program for supervisors. He is sure it is successful but he has no measured results. In discussing it he said, "If I had only thought to count the number

of frowns I got in the department before the training, and the number of smiles I get now, maybe the data would be statistically valid at at least the 5 per cent level." At heart we are still strictly scientific. Practice has forced us to make decisions on bases which cannot be scientifically proven. We have learned that a workable solution on time is worth more than a perfect one too late, so we don't publish.

We are not alone in this dilemma. Dr. Cameron of the National Industrial Health Service, at the meeting of Psychiatrists and Psychologists at Asbury Park told us of many health projects set up by industry, and pleaded with us for some way of measuring their success. We cannot usually in industry set up experimental controls. For example, Dr. Cameron in talking to me said, "You have a visiting nurse? Does she pay for herself?" I am sure she does but I can't prove it. Of course under laboratory conditions we could set up controls; we could give nurses' services to half of our office force and not to the other half. Then we could keep track of absenteeism, turnover and even make morale surveys for the two halves, and perhaps come out with proof but can you see any company being willing to set up such a program? I certainly cannot see myself asking, much less advising the Aetna to go to any such measures to prove something, which we already think we know.

Perhaps we are still too conscious of our heritage that any idea to be worth publishing must represent research and valid proof. Perhaps knowing the complications of human behavior we become so involved when we try to write in general terms and for popular consumption we put in so many "ifs and buts" that we get discouraged and leave the writing to the nonpsychologists who can go all out for a given plan and forget the complications of which we are so conscious. Perhaps since we do not need to publish to advance in our work, and since we are fairly busy, we get a little lazy and do not take the time and energy to clarify our thinking and put it down on paper for others to read

and maybe profit by. But the fact remains, —we do not publish.

In talking with our Medical Department they tell me that they have two types of journals—one of a strictly scientific nature where 80 per cent at least of the contributions come from research centers, and another type where the contributions are mostly from practicing physicians, and maybe reports of single cases, or small groups of cases. No one expects them to be valid research articles but they are often very suggestive.

Is this our solution? Perhaps, but before it can work we must change some of our psychological thinking. It's a very thin and sometimes wavy line between the fundamental concepts that form at least a part of our contribution to management and our acceptance of the less rigid principles of proof that prevail in the management field. Would our publications in this less rigid field add anything to the fundamental knowledge of psychology? Perhaps the trees are so thick that we do not see the forest and perhaps we have got to wait until some of our colleagues who have been through the experience retire, and getting at a distance, which gives them an objective viewpoint, become our spokesmen. Perhaps our real function is that of a liaison officer between our experimental workers and management under which function our chief duty would be to keep very well informed on both sides, and display the ingenuity to connect them, even when in many cases the connection is far from obvious. I know of one case where a strictly experimental study by Berth and Rabinowitz on the two cord problem helped to set up a change in a sales training course for salesmen.

I realize that this talk has been full of "perhaps," which means that questions have been raised, and no conclusions reached, but psychologists in private industry are only about 100 strong and we need the advice of our consulting friends and especially of our academic ones to help us see clearly where our greatest contribution to a young but fast growing science lies.

## Calling in Psychologists Early

A few years ago, military men found that a lot of new weapons were getting too complicated for the men who had to operate them. Industry found much the same thing with new plant equipment. Electronic and mechanical devices had been added so fast that the human mind could not keep up. Applied psychologists were called in to "humanize" the machines.

Shortly after the end of World War II, the psychologists were turned loose on the nearly complete designs for new machines to be used for military production. At that point, about all the psychologists could do was change dials for easier reading, color or illumination for less eye strain, size or shape of knobs and wheels for easier identification, and a few other minor things that would not hurt

the over-all engineering. Obviously this helped some, but it was not enough. Private industry was even slower to tackle the problem, largely because of the expense involved in re-designing machines.

To remedy this, consultant firms, such as Dunlap & Associates, Inc., of Stamford, Conn., are campaigning for a place in the early design stages of equipment development. Most government groups favor this new approach, but industry, in general, is skeptical. Industry seems to agree that more consideration should be given to the human factors early in the design process. But it does not think that design engineers are going to be happy about psychologists butting into the blueprint phase of the problem. (*Business Week*, December 20, 1952.)

## Book Reviews

Wolfe, D., Buxton, C. E., Cofer, C. N., Gustad, J. W., MacLeod, R. B., and McKeachie, W. T. *Improving undergraduate instruction in psychology*. New York: Macmillan, 1952. Pp. vii + 60. \$1.10.

Surely it is now even more true than when H. G. Wells made the statement, that there is a race between education and catastrophe. The committee making this report is able, headed by a man who was probably closer than any other to the work of psychologists during the war and in the immediate post-war period. From this group might therefore be expected a program having as a major feature, a broad and stimulating view of the vital importance of psychology in the present-day world.

But national and world affairs seem not even mentioned in the volume. One would never know there had been a world war! The first chapter, on objectives, emphasizes "the contribution which psychology can make to a liberal college education," but the concept of such education seems formal and remote from the current scene. Nor should the major objective of first work in psychology be to foster students' "personal growth and increased ability to meet personal and social adjustment problems adequately." A four-page chapter on "Personal Adjustment Courses" declares scornfully that "it is no more justified to consider such a course as a course in psychology than it would be to substitute . . . a course on household repairs for introductory physics" (p. 41). And courses which "deal with special interest areas or purport to provide technical training" are only tolerated; there is admiration for "a few conscientious departments, determined to provide the best possible training for students, which have recommended that such courses be eliminated, even at the risk of decreasing enrollments and displeasing other departments" (p. 24). The major chapter, on "The Recommended Curriculum," urges a first course giving "a systematic presentation of scientific content" followed by core courses on "motivation, perception, thinking and language, ability," and advanced courses in social psychology, physiological psychology, etc. A two and a half

page chapter on "Technical Training in Psychology" suggests that after such an undergraduate program, "a few months of full-time vocationally-oriented training in a post-A.B. institute could give the student a battery of job skills" (p. 44). And the concept of "liberal college education" becomes fairly clear: a program in which psychology need feel no responsibility for world or national or community problems, or student welfare, or vocation—and may smugly go its own self-centered way.

A six-page chapter on "Implementation of the Curriculum" points out (for instance) that, though the proposed program may bring some reduction in number of courses, staff can be absorbed by laboratories. A final brief chapter on "Research Problems underlying the Curriculum" suggests (for example) appraisal of the first course by number of students taking further work in psychology, and touches briefly on methods of instruction; in spite of its title, the volume deals with this last topic only incidentally—is given over to emphatic declaration for a systematic theoretical undergraduate program, and impatient belittling of alternatives.

Such a partisan position can be adequately appraised only by a balancing comparison with alternatives. Presumably an alternative report might emphasize that indeed "wars begin in the minds of men," that psychological warfare is more powerful than the H-bomb, that psycho-social problems are major in any nation and any community—and that psychologists should courageously do anything they can to bring general understanding of these issues. It might be proudly confident that psychology had much to offer students in better understanding themselves and their problems, and that such help could be a vital part of a broad systematic treatment of the subject. It might be exhilarated by the vocational usefulness of much psychological material, and find therein enrichment of its essential subject-matter. Instead of a statement which many would have considered conservative and professionally introvert thirty years ago, it might be a document which would give college administrators and faculty in other departments a stimu-

lating view of psychology as a science rapidly advancing and eager to cooperate in efforts to build educational programs more fully meeting the problems of the present world. The one-sidedness and inadequacy of the present little volume seems to the reviewer to emphasize a need for such an alternative document. Lacking it, the hope must be that the Education and Training Board may take a position more positive and forward-looking.

Sidney L. Pressey

Ohio State University

Hirsh, I. J. *The measurement of hearing*. New York: McGraw-Hill Book Co., 1952. Pp. ix + 364. \$6.00.

This book is concerned with the information about acoustics, electro-acoustic equipment, psychology of hearing and related topics that is basic to both the clinical and experimental measurement of various aspects of hearing. Written by an experimental psychologist thoroughly acquainted with psychophysical methods, the treatise stresses psychological contributions. It is designed for use as a reference by those engaged in measuring and treating hearing disorders, as a text for those preparing for this kind of clinical work, and as reference material for all those interested in hearing.

An introduction to psychophysical measurement is followed by discussion of the principles of sound and electricity basic to control, production and measurement of auditory stimuli. These principles are then applied to operation of electro-acoustic equipment. Various kinds of auditory measurement used in clinical audiometry are discussed in relation to clinical procedures. Each section on auditory measurement is followed by clinical applications and the last chapter is devoted entirely to clinical audiometry.

Applied psychologists will be interested mainly in the sections devoted to clinical applications. Nevertheless, a clear understanding of the principles underlying sound clinical practice is possible only if the sections on experimental findings are consulted.

Although the author has done an excellent job of organization and clear exposition with highly technical subject matter, the reader

must not expect to absorb the material without concentrated study. This excellent book is a must for anyone preparing to do research in the measurement of hearing as well as for all those with a serious interest in the field. Readers will be especially grateful for the completeness of the information that accompanies the figures and for the glossary of technical terms. It is probable that the book will find its greatest use as a reference work on techniques of measurement and clinical applications.

Miles A. Tinker

University of Minnesota

Campbell, C. M. (Editor). *Practical applications of democratic administration*. New York: Harper & Brothers Publishers, 1952. Pp. 325. \$3.00.

Generally speaking school administrators in the United States have voiced their allegiance to the broad concepts of democracy. How these concepts can be given vehicle in schools and school administration is far from clear to many superintendents and principals.

*Practical Applications of Democratic Administration* represents in condensed form the thinking of a dozen scholars on this question. Their contribution is predicated upon a sound philosophical basis, which turns early to practical applications based upon research and experience.

Leadership in educational administration is the ribbon which binds the separate contributions together in a package which should be both appealing to, and much sought for by, all professional educators. The two chapters dealing with sociology and psychology are strikingly illustrative of the integration of broad fields of understanding which democratic leadership must draw upon. These same two chapters offer to many present-day school administrators a challenge for considering more realistically than they have been accustomed to, the social forces in their school communities and the vestiges of autocracy which individuals and groups have inherited.

Readers are not left with a "so what" attitude, because seven chapters follow immediately and describe in clear expository style actual administrative leadership practices in a number of school communities.

At times it is somewhat difficult to align the separate examples with specific prongs of the foregoing theory. This is, however, understandable because democracy does not propose to "blueprint" practice. Instead the stream of democracy is fed by leadership from local tributaries each carrying in suspension particles unique to its own fields of origin.

The role of administrative leadership is not easy as pointed out in the concluding chapters, yet the present volume is filled with the necessary materials. Although democratic idealism seems to have swept the country, the next step suggested for educators is to reach consensus on somewhat more specific points of both theory and practice. It is unlikely that this can be accomplished in a setting which is pessimistic.

The implication is clear that a science of human engineering coupled with educational statesmanship has germinated and is in serious need of cultivation. Neglect in this area points to a spotty blighted harvest that will represent only a fractional part of the potential. Therefore, carefully formulated experimental programs like those cited in this brief volume should become the usual rather than the unusual practice of current and future educational leadership in our democratic society. To do this will take school superintendents and principals away from many of their present routines and "behind the desk" management activities out into the community. This will not sell the school short, however, because the community will bring back to it a richness otherwise unattainable.

Hugh M. Shafer.

*School of Education,  
University of Pennsylvania*

Dooher, M. J., and Marquis, Vivienne (Eds.). *The development of executive talent*. New York: American Management Association, 1952. Pp. 576. \$6.75. (\$5.75 to AMA members.)

For anyone who wants to develop a centralized planned economy in the United States, this book will offer little of value for it is oriented to: "Management's role in the preservation of a free society is putting the real meaning of a free society to work within the organization for which each individual

executive is responsible. The basic objective is the development of individuals. . . . The basic purpose of management is absolutely consistent with that of a free society, and the individual manager's responsibility is to work that way."

For anyone who has negative reactions to "management-minded" research and publications or who feels that business management as a function in our society is part of our reactionary past, this book will offer little of interest. For this book is based upon the principle of "management as a profession, a science, and an art." It recognizes that management has an aggressive, dynamic role to play in the major national and international struggle between two ways of life, but that the successful performance of this role is greatly dependent upon the development of capable leadership. The purpose of the book, then, is to bring together the productive experiences and practices of many organizations in their efforts to produce management and executive personnel who will function most effectively in achieving the goals of a free society.

From this it should not be inferred that the book is either political or controversial. The main body of the book is divided into nine parts, consisting of 50 chapters, contributed by 44 authors from business and educational organizations. The subjects covered include: setting up the program—basic principles and practices; organization planning; putting the program into action; conference training methods; special approaches, techniques and programs; getting results from follow-up counseling; program evaluation; and trends in management development. The remaining pages, consisting of about one-half of the entire volume, are devoted to case study reports of methods used by such companies as Standard Oil (N. J.); United Parcel Service; Sears, Roebuck and Co.; Detroit Edison; U. S. Rubber; Westinghouse; and others. In addition, there is an extensive bibliography of approximately 400 items divided under a variety of sub-topics relating to leadership and management.

Most of the chapters offer a combination of practical "how to do it" material and discussions from the experimental literature. Although some academicians may be disap-

pointed in a large part of the material, it does present a carefully planned compromise approach for the practicing businessman and the classroom educator. While no step-by-step solution to individual problems is offered, a pattern of action is noted. In the final analysis, the book contains specific, practical guidance on all of the problems involved at every stage of planning and administration—from the analysis of needs, through the discovery of latent executive ability, to the inventorying, rating and development of executive skills.

C. G. Browne

Wayne University

Judd, D. B. *Color in business, science, and industry*. New York: John Wiley & Sons, Inc., 1952. Pp. 401. \$6.50.

The measurement and specification of color have undergone great advances during the past thirty years. This book should prove a fruitful venture because, during the period of maximum development in colorimetry, the author has become an outstanding authority and has held a strategic position at the National Bureau of Standards. The scope of the book is broad and records everything that has appealed to the writer as pertinent or interesting in respect to color. Such a treatment cannot be expected to be complete since the thinking of the author is set down as final without consideration of alternative facts and theories.

The background of the book is physical both in the material presented and in the point of view. Other influences have made themselves felt, but have not altered the treatment. Psychology, for example, is mentioned frequently, but one discovers that it refers to "the customer's angle" rather than to available scientific material. Much is said about psychophysics, but it is not the classical psychophysics of Fechner, Müller and Titchener. Methodological considerations of the operations necessary for adequate observation are not involved. An implied definition is that visual psychophysics is a study of radiation modified by the sensitivity of the eye.

The presentation is in three parts without further conventional subdivision into chapters. Part I is a compilation of "basic facts"

which have entered into the author's thinking on the subject. The treatment is unsystematic. Materials of physiological, physical and psychological character are intermingled. No distinction is made between data and hypotheses and the philosophy is that of naïve physical realism. Nevertheless, the exposition proceeds from physiology of the eye to the tristimulus mixture of colors including radiation by the way. It is difficult to assess the pertinence of various topics. The reader is forced into an item by item evaluation which must be confusing to a novice in the field. It would have been sufficient for the remainder of the book had the author limited the introduction to a statement of the tristimulus hypothesis and the development of its colorimetric implications. As the discussion stands, it is not clear precisely what the author himself holds to be the "three color hypothesis." He states that "we have seen that normal color vision is tridimensional" (p. 67), but whether this derives from "the fact that we get three independent kinds of information from the cones, light-dark, red-green, yellow-blue" (p. 18) or from the hypothesis that "some of the cones contain short-wave absorbing (V) pigment, some contain a preponderance of long-wave absorbing (R) pigment, and some contain a preponderance of middle-wave absorbing (G) pigment" (p. 18) is not indicated.

In Part II we have the meat of the book under the title "Tools and Technics" of colorimetry. It occupies some two-thirds of the text and gives precise and comprehensive information needed to carry out measurements of color and to specify them in one of the alternative systems of notation. The reviewer considers this work one of the outstanding presentations of colorimetry, one that may very well become the standard reference in the field.

It is regrettable that the author felt the necessity to go beyond this field. Business and industry will hardly be concerned with the physiological hypotheses of vision nor the psychophysical techniques of psychology. Moreover, no one can expect to speak with authority on all subjects.

Forrest L. Dimmick

U. S. Naval Medical Research Laboratory,  
New London, Connecticut

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Problems of consciousness.* H. A. Abramson, Editor. The Josiah Macy, Jr. Foundation. New York: Corlies, Macy and Co., 1952. Pp. 153. \$3.25.
- Geography of living things.* M. S. Anderson. New York: Philosophical Library, 1952. Pp. 202. \$2.75.
- Statistical theory in research.* R. L. Anderson and T. A. Bancroft. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 399. \$7.00.
- Human relations and administration.* Kenneth R. Andrews, editor. Cambridge: Harvard University Press, 1952. Pp. 271. \$4.50.
- An introduction to field theory and interaction theory.* Chris Argyris. New Haven: Labor and Management Center, Yale University, 1952. Pp. 71.
- Design for a brain.* W. Ross Ashby. New York: John Wiley and Sons, Inc., 1952. Pp. 259. \$6.00.
- Mental prodigies.* Fred Barlow. New York: Philosophical Library, 1952. Pp. 256. \$4.75.
- Operations research.* James H. Batchelor. Cleveland: Case Institute of Technology, 1952. Pp. 95. \$1.00.
- How to write advertising that sells.* Clyde Bedell. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 519. \$6.00.
- A manual for the differential aptitude tests.* George K. Bennett, Harold G. Seashore, and Alexander G. Wesman. New York: The Psychological Corporation, 1952. Pp. 77.
- Mental hygiene for classroom teachers.* Harold W. Bernard. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 472. \$4.75.
- Practical psychology.* Karl S. Bernhardt. New York: McGraw-Hill Book Co., Inc., 1953.
- The infirmities of genius.* W. R. Bett. New York: Philosophical Library, 1952. Pp. 192. \$4.75.
- The philosophy of social work.* Herbert Bisno. Washington, D. C.: Public Affairs Press, 1952. Pp. 139. \$3.25.
- Research in the training of teachers.* Henry Bowers. Toronto: J. M. Dent and Sons, Ltd., 1952. Pp. 167. \$1.90.
- General psychology.* Revised edition. Robert Edward Brennan. New York: The Macmillan Co., 1952. Pp. 524. \$5.50.
- Applied psychology.* Harold Ernest Burt. New York: Prentice-Hall, Inc., 1952. Pp. 480.
- The psychology of learning.* James Deese. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 398. \$5.50.
- An outline of nonacademic personnel administration in higher education.* Donald E. Dickason. Champaign: Donald E. Dickason, 809 S. Wright St., 1952. Pp. 52. \$2.00.
- The dynamics of social action.* Seba Eldridge. Washington, D. C.: Public Affairs Press, 1952. Pp. 115. \$2.50.
- The scientific study of personality.* H. J. Eysenck. New York: The Macmillan Co., 1952. Pp. 298. \$4.50.
- Deaf children in a hearing world.* Miriam Forster Fiedler. New York: The Ronald Press Co., 1952. Pp. 320. \$5.00.
- Let's hear it!* George W. Frankel. New York: Stratford House, Inc., 1952. Pp. 63. \$1.00.
- Understanding old age.* Jeanne G. Gilbert. New York: The Ronald Press Co., 1952. Pp. 442. \$5.00.
- Problems of the family.* Fowler V. Harper. Indianapolis: Bobbs-Merrill Co., Inc., 1952. Pp. 850. \$9.00.
- Appraising personality.* Molly Harrower. New York: W. W. Norton and Co., Inc., 1952. Pp. 197. \$4.00.
- The fundamentals of social psychology.* Eugene L. Hartley and Ruth E. Hartley. New York: Alfred A. Knopf, Inc., 1952. Pp. 832. \$5.50.
- Psychoanalysis as science.* Ernest R. Hilgard, Lawrence S. Kubie, and E. Pumpian-Mindlin. Stanford, Calif.: Stanford University Press, 1952. Pp. 158. \$4.25.
- The measurement of hearing.* Ira J. Hirsh. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 364. \$6.00.
- The treatment of the young delinquent.* J. Arthur Hoyles. New York: The Philosophical Library, 1952. Pp. 261. \$4.75.
- Some principals of construction of group intelligence tests for adults.* Husen and Henrickson. Stockholm: Almqvist and Wiksell, 1951. Pp. 98.
- The Ames demonstrations in perception.* William H. Ittelson. Princeton: Princeton University Press, 1952. Pp. 81. \$4.00.
- Measurement in education.* Arthur M. Jordan. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 521. \$5.25.
- The Tree test.* Charles Koch. New York: Grune and Stratton, Inc., 1952. Pp. 87.
- Evolution and human destiny.* Fred Kohler. New York: The Philosophical Library, 1952. Pp. 118. \$2.75.
- Functional neuroanatomy.* Wendell Krieg. New York: The Blakiston Co., 1953. Pp. 645. \$9.00.
- Psychological studies of human development.* Raymond G. Kuhlen and George G. Thompson. New York: Appleton-Century-Crofts, Inc., 1952. Pp. 533. \$3.50.
- A history of psychology in autobiography.* Vol. IV. Herbert S. Langfeld et al., editors. Worcester: Clark University Press, 1952. \$7.50.
- How to talk with people: a guide for the improvement of communication in committees.* Irving J. Lee. New York: Harper and Brothers, 1952.

- Prescription for rebellion.* Robert Lindner. New York: Rinehart and Co., Inc., 1952. Pp. 305. \$3.50.
- Powers of the mind.* Paul Maslow. Vol. II of The Life Science. Brooklyn: Paul Maslow, 16 Court Street, 1952. Pp. 153. \$3.50.
- Man's search for himself.* Rollo May. New York: W. W. Norton and Co., 1953. Pp. 277. \$3.50.
- Finality and form.* Warren S. McCulloch. Springfield: Charles C Thomas, Publisher, 1952. Pp. 63. \$3.75.
- The focussed interview.* Robert K. Merton, Marjorie Fiske, and Patricia Kendall. New York: Bureau of Applied Social Research, Columbia University, 1952. Pp. 202. \$3.00.
- Social and psychological factors in opiate addiction.* Alan S. Meyer, editor. New York: Bureau of Applied Social Research, Columbia University, 1952. Pp. 169. \$1.00.
- Office psychiatry.* L. G. Moench. Chicago: Year Book Publishers, Inc., 1952. Pp. 299. \$6.00.
- The philosophy of psychiatry.* Harold Palmer. New York: Philosophical Library, 1952. Pp. 63. \$2.75.
- The sensations, their functions, processes, and mechanisms.* Henri Piéron. New Haven: Yale University Press, 1952. Pp. 469. \$6.00.
- The inmates.* John Cowper Powys. New York: Philosophical Library, 1952. Pp. 318. \$4.50.
- Advanced statistical methods in biometric research.* C. Radhakrishna Rao. New York: John Wiley and Sons, Inc., 1952. Pp. 390. \$7.50.
- The secret self.* Theodor Reik. New York: Farrar, Straus and Young, Inc., 1952. Pp. 329. \$3.50.
- Volunteer work camp.* Henry W. Riecken. Cambridge: Addison-Wesley Press, Inc., 1952. Pp. 260. \$3.60.
- Administering changes: a case study of human relations in a factory.* Harriet O. Ronken and Paul R. Lawrence. Boston: Division of Research, Harvard Business School, 1952. Pp. 324. \$3.50.
- Medical public relations.* Schuler, Mowitz, and Mayer. Ann Arbor: Edwards Brothers, Inc., 1952. Pp. 227.
- Fundamental concepts in clinical psychology.* G. Wilson Shaffer and Richard S. Lazarus. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 493. \$6.00.
- A practical guide for troubled people.* Lee R. Steiner. New York: Greenberg, Publisher, 1952. Pp. 299. \$3.50.
- Introduction to logical theory.* P. F. Strawson. New York: John Wiley and Sons, Inc., 1952. Pp. 263. \$3.50.
- Experimental diagnostics of drives.* Med. L. Szondi. New York: Grune and Stratton, Inc., 1952. Pp. 272. \$13.50.
- Our common neurosis.* Charles B. Thompson and Alfreda P. Sill. New York: Exposition Press, 1953. Pp. 210. \$3.50.
- Introduction to testing.* Arthur E. Traxler et al. New York: Harper and Brothers, 1952. Pp. 394. \$5.00.
- Out of step.* Joseph Trenaman. New York: The Philosophical Library, 1952. Pp. 217. \$4.75.
- Red Wing—five years later.* Roland S. Vaile. Minneapolis: University of Minnesota Press, 1952. Pp. 27. Gratis.
- A further study of visual perception.* M. D. Vernon. New York: Cambridge University Press, 1952. Pp. 263. \$7.00.
- Statistical tables and problems.* (Third edition.) Albert Waugh. New York: McGraw-Hill Book Co., Inc., 1952. Pp. 242. \$3.00.
- Range of human capacities.* Second edition. David Wechsler. Baltimore: The Williams and Wilkins Co., 1952. Pp. 190. \$4.00.
- Contributions toward medical psychology.* Arthur Weider, editor. New York: The Ronald Press Co., 1952. Pp. 885. \$12.00.
- Personnel interviewing.* James D. Weinland and Margaret V. Gross. New York: Ronald Press Co., 1952. Pp. 416. \$6.00.
- Improving undergraduate instruction in psychology.* Dael Wolfle, et al. New York: The Macmillan Co., 1952. Pp. 60. \$1.25.
- Personality and problems of adjustment.* Kimball Young. New York: Appleton-Century-Crofts, Inc., 1952. Pp. 716. \$5.00.
- Helping parents understand the exceptional child.* Child Research Clinic. Langhorne, Pa.: The Woods Schools, 1952. Pp. 42. Available on request.
- Selection, training, and use of personnel in industrial research.* Proceedings of the Second Annual Conference on Industrial Research. New York: King's Crown Press, 1952. Pp. 274. \$4.50.
- Employee personnel practices in colleges and universities—1951-1952.* Champaign: College and University Personnel Association, 1952. Pp. 69. \$2.50.

# Journal of Applied Psychology

Vol. 37, No. 3

JUNE, 1953

## The Measurement of Leadership Attitudes in Industry

Edwin A. Fleishman \*

*USAF Air Training Command, Human Resources Research Center \*\**

Recent years have seen an intensified concern in industry for the problems of leadership and human relations. Evidence of this can be seen in the increasing number of leadership training programs which have been instituted in various industries. However, those who train supervisors must still rely on a limited number of general assumptions largely unsupported by either sound theory or empirical data. Part of this difficulty arises from the fact that effectual leadership depends to a great extent on the situation. Additional difficulties stem from the lack of adequate criteria of group effectiveness. A pressing need is the development of dependable research instruments which can be utilized to describe adequately the various complex socio-psychological aspects of a wide variety of leader-group situations.<sup>1</sup> If these were available, they might later be related to criteria of group effectiveness in many specific situations in which leaders function. The present study was a further attempt to develop a number of such instruments which would have application in industry.

\* This research was carried out while the writer was at the Personnel Research Board, Ohio State University, as part of a larger project on leadership in industry, with the cooperation of the International Harvester Company.

\*\* Lackland Air Force Base, San Antonio, Texas. The opinions or conclusions contained in this report are those of the author. They are not to be construed as reflecting the views or indorsement of the Department of the Air Force.

<sup>1</sup> The Personnel Research Board has made the development of such instruments a major part of their leadership research program. See Stogdill and Shartle (10), Shartle (9), Seeman (8), Halpin and Winer (4), Hemphill (5), Hemphill and Westie (6), and Fleishman (1, 2). Another approach to the measurement of leadership attitudes has been made by Nelson (7).

In a previous paper (2) the writer has described a questionnaire found useful for the description of supervisory *behavior*. The present paper describes questionnaires which were developed for the measurement of leadership *attitudes*.

### Construction of the Questionnaire

A preliminary 110-item Leadership Opinion Questionnaire was administered to 100 foremen in a pilot study at the company's Central School. These foremen represented 17 different company plants. The foreman indicated for each item how frequently he thought he should do what each item described. He responded by marking one of five frequency alternatives which followed each item (e.g., always, often, occasionally, seldom, never). He was told that there were no right or wrong answers in the questionnaire since "everyone's work group is different and what is the best way to lead one group may not be the best way for another."

The items in this questionnaire were generally parallel to those in the pre-test form of the Supervisory Behavior Description previously described (2). However, in this latter questionnaire the items were worded in terms of "what does your own supervisor actually do" while in the present questionnaire items were worded in terms of "what *should* you do." The questionnaire was scored along two major and two minor dimensions.<sup>2</sup> Of the

<sup>2</sup> These dimensions were originally isolated in a factor analysis of the items of a Leadership Behavior Description questionnaire administered to 300 Air Force crew members who described their airplane commander (4). Later analysis of the items, based on this industrial population, supported only the two major factors "Consideration" and "Initiating Structure" (2).

two major dimensions, one was called "Consideration," which contained items reflecting the extent to which the supervisor is considerate of the feelings of those under him. It comes closest to representing the "human relations" approach toward group members. The second major dimension was called "Initiating Structure," and contained items reflecting the extent to which the supervisor facilitates or defines group interactions toward *goal attainment*. He does this by planning, communicating, scheduling, criticizing, initiating new ideas, etc. The two minor factors were called, "Production Emphasis," and "Social Sensitivity." Response distributions were obtained for the alternatives to each of the items in the questionnaire.

The corrected split-half reliability estimates for the two major keys "Consideration" and "Initiating Structure" were .69 and .73, respectively, and for the two minor keys "Production Emphasis" and "Social Sensitivity" the reliabilities were .36 and .33, respectively. In the light of the low reliabilities of the latter two keys and in view of the fact that a modified factor analysis of the items in the parallel Supervisory Behavior Description indicated that only the two major dimensions were meaningful in this industrial population, the dimensions of "Production Emphasis" and "Social Sensitivity" were omitted from the revised form.<sup>3</sup>

The criteria for selecting items for the revised form included: (1) the response distributions of the items in the Leadership Opinion Questionnaire; and (2) the factor loadings, based on this industrial population, of parallel items on the Supervisory Behavior Description. Items were favored whose parallel item had a high loading on the factor in which it was keyed and insignificant loading on the other factor. It was hoped that this procedure would yield two scales tapping relatively independent leadership attitude dimensions.

Twenty items were selected in this manner for the "Consideration" key and 20 items

were selected for the "Initiating Structure" key. Examples of items in the revised "Consideration" key were:

- Help people in the work group with their personal problems.
- Back up what people under you do.
- Speak in a manner not to be questioned (response weights reversed).

Examples of items in the revised "Initiating Structure" key were:

- Emphasize meeting of deadlines.
- Assign people in the work group to particular tasks.
- Meet with the work group at regularly scheduled times.

#### Administration of the Revised Questionnaires

Various forms of the revised questionnaire were administered in one of the company's plants.

A total of 122 foremen filled out the following forms with the indicated response "sets":

(1) A *Leadership Opinion Questionnaire*: How he thinks he should lead his own work group.

(2) A questionnaire entitled, "*What Your Boss Expects of You*": A description of how the foreman feels his boss wants him to lead the work group.

A total of 394 workers filled out a questionnaire entitled, "How You Expect an Ideal Foreman to Act." This is a description of worker expectations regarding leadership behavior.

Also, 60 supervisors above the rank of foreman filled out the following questionnaires:

(1) A *Leadership Opinion Questionnaire*: How the boss thinks he should lead the foremen under him.

(2) A questionnaire entitled, "*What You Expect of Your Foremen*": A description by the boss of how he wants his foremen to lead their workers.

All these forms are variations of the Leadership Opinion Questionnaire revised on the basis of the pilot study. All contained the same 40 items reworded slightly in certain forms to apply to the appropriate situational context.

<sup>3</sup> Actually, in this analysis it appeared that in the industrial sample, "Initiating Structure" and "Production Emphasis" were reflections of the same underlying dimension, as were "Consideration" and "Social Sensitivity."

Table 1  
Means, Range, Standard Deviations, Reliabilities, and Intercorrelations of  
Dimension Scores in Each Revised Instrument

Instrument	Dimension <sup>1</sup>	Mean Score	S.D.	Range <sup>2</sup>	Reliability Estimate <sup>3</sup>	Inter-correlation
<i>Filled out by foremen (N = 122)</i>						
Leadership Opinion Questionnaire	Consideration	53.9	7.2	36 to 74	.70	-.01
	Initiating Structure	53.3	7.8	34 to 69	.79	
"What Your Boss Expects of You"	Consideration	48.5	10.2	21 to 68	.87	.05
	Initiating Structure	51.2	8.4	31 to 68	.78	
<i>Filled out by workers (N = 394)</i>						
"How You Expect an Ideal Foreman to Act"	Consideration	57.0	5.5	41 to 70	.89	.04
	Initiating Structure	44.2	3.9	26 to 58	.88	
<i>Filled out by foreman's boss (N = 60)</i>						
"What You Expect of Your Foremen"	Consideration	53.0	7.3	38 to 67	.64	-.31
	Initiating Structure	54.0	6.7	37 to 68	.78	
Leadership Opinion Questionnaire	Consideration	58.0	6.4	40 to 75	.60	-.23
	Initiating Structure	52.4	7.6	31 to 69	.82	

<sup>1</sup> Each dimension key in each questionnaire contained 20 items.

<sup>2</sup> Since alternatives to each item were weighted zero to four, the highest possible score is 80 in each questionnaire for each dimension.

<sup>3</sup> Split-half correlations corrected to full length of each dimension by the Spearman-Brown formula.

### Results

*Adequacy of the Questionnaires.* On all the instruments, the five alternative responses to each item were assigned weights from zero to four. Whether the high frequency alternative (e.g., always) was weighted zero or four depended on the item's orientation with respect to the total dimension continuum. Total dimension scores were derived by adding the weights corresponding to the alternatives marked for the items in each dimension. Table 1 presents a summary of the means, range of scores, standard deviations, reliabilities and dimension intercorrelations for each instrument.

The striking feature of Table 1 is the independence of the two dimensions in each of the forms used. This is especially true when the forms are filled out by workers and by foremen. The correlations cluster about zero. Even in the case of the foremen's supervisors, these correlations are low relative to those usually obtained with such instruments and do not reach the 1% level of significance.

The important thing in interpreting the reliability coefficients is their magnitude *relative* to the dimension intercorrelations. Apparently, these instruments tap reliably two *independent* dimensions of leadership attitudes. This is especially interesting since a criterion for item inclusion was the loading of a *parallel item* in a Supervisory Behavior Description questionnaire. An ideal but time consuming procedure would have been to repeat the factor analysis on the *attitude* form but the independence of dimensions seems to have been accomplished by the present procedure. At least it appears that the usual "halo" effect, which often inflates the intercorrelation among keys in instruments of this type, has been efficiently partialled out in the revised form. The distributions of scores obtained from most of the questionnaires are roughly normal in shape.

The implication of these findings seems to be that the dimensions of "Consideration" and "Initiating Structure" are as meaningful and as independent in the attitudinal domain

Table 2

Comparison of the Leadership Attitude Scores of  
Workers, Foremen, General Foremen,  
and Superintendents

Dimension	Level in the Organization	Mean	S.D.
"Consideration"	Superintendents (N = 13)	52.6	8.1
	General Foremen (N = 30)	53.2	7.1
	Foremen (N = 122)	53.9	7.2
	Workers (N = 394)	57.0 <sup>1</sup>	5.5
"Initiating Structure"	Superintendents (N = 13)	55.5	5.7
	General Foremen (N = 30)	53.6	6.9
	Foremen (N = 122)	53.3	7.8
	Workers (N = 394)	44.2 <sup>1</sup>	3.9

<sup>1</sup> Indicates this mean differs significantly (beyond the .01 level of confidence) from the mean of the foremen group.

of leadership as in the behavioral realm. It thus appears that supervisors may be high in the amount of consideration they feel should be shown their subordinates, but at the same time may be either low or high in the amount of planning, criticizing, pushing for production, and general "structuring" behavior that they feel they should engage in. There is also the indication that workers who want a great deal of "consideration" in their foremen do not necessarily want less "structuring" or more "structuring" of their work activities from him.

*Attitudes at Different Levels.* The questionnaires entitled "What You Expect of Your Foremen" (filled out by supervisors), Leadership Opinion Questionnaire (filled out by foremen), and "How You Expect an Ideal Foreman to Act" (filled out by workers) all measure the respondents' values about how the *work groups* should be led. The mean scores on each instrument provide a comparison of these leadership attitudes at four levels in the plant. Table 2 presents this comparison for four clear-cut organizational levels.

The comparison shows that the attitudes of the foreman group are much more like the attitudes of superiors than they are like the attitudes of the workers. Differences between the mean scores of the foremen and their bosses are not statistically significant. Differences between the scores of the foremen and those of the workers are highly significant. This is true of scores on both leadership dimensions. The workers prefer more "consideration" and less "structure."<sup>4</sup> It also appears that the higher people were in the plant hierarchy, the less "consideration" they felt the workers should get. Moreover, the higher the level, the more "structuring" the people felt should be initiated with the work group. However, some of these differences were not large or significant although consistent. The tendency was for the foremen's attitudes to fall somewhere between what the workers expect and what their supervisors expect.

Table 2 also indicates the relatively small standard deviations of the scores made by workers on both dimensions of the form "How You Expect an Ideal Foreman to Act." In each dimension the differences between the sigmas of worker attitude scores and that for supervisor attitude scores are statistically significant ( $P < .01$ ). It appears that the workers are more homogeneous with respect to their leadership expectations than are the supervisory groups with respect to how they feel groups should be led. However, these scores present little evidence revealing the existence of an "ideal leadership" stereotype among workers since there was still a considerable range of scores on both expected "consideration" and expected "structure" at the worker level (see Table 1).

It was also possible to compare the leadership attitudes of supervisors above the rank of foreman toward foremen and workers. This comparison is between scores made by these supervisors on the Leadership Opinion Questionnaire and the form "What You Expect of Your Foremen." This comparison is presented in Table 3.

<sup>4</sup> It is interesting to note that although the workers are generally on a piece rate basis, they prefer less "structuring," which consists in large part of foremen activities pushing for production.

Table 3

Comparison of the Leadership Attitudes of Foremen's Supervisors Toward Workers and Foremen (N = 60 Supervisors)

Dimension	Leadership Attitudes				P	r <sub>12</sub>
	Toward Workers		Toward Foremen			
	Mean	S.D.	Mean	S.D.		
"Consideration"	53.0	7.3	58.0	6.4	P < .01	.58
"Initiating Structure"	54.0	6.7	52.4	7.6	P > .05	.73

It can be seen that supervisors above the rank of foreman scored significantly higher on their "Consideration" attitudes toward the foremen than in their "Consideration" attitudes toward workers. However, the difference in the amount of "Structuring" they felt should be initiated toward each group is not statistically significant. Moreover, bosses who scored high in their "Consideration" attitudes toward foremen tended also to score higher on these attitudes toward workers ( $r = .58$ ). This was also true for "Initiating Structure" attitudes toward foremen and worker groups ( $r = .73$ ).

*Differences between Work Groups in Their Leadership Expectations.* An analysis of variance was made of the scores derived from 226 workers, drawn from 73 different work groups on the questionnaire "How You Expect an Ideal Foreman to Act." This analysis revealed significant differences between work groups relative to that within work groups ( $F = 14.7$ ,  $P < .01$ ) in how "considerate" they expect an ideal foreman to be. Apparently, worker attitudes concerning the amount of "consideration" desired depends to a large extent on the particular work groups. However, differences between work groups in how much "structuring" behavior they felt the foremen should engage in were not significant. This may be due to the small variation in scores on this dimension for the total sample of workers ( $\sigma = 3.9$ , see Table 2).

*Relationships with Labor Grievance Rates.* A problem of future research with these instruments is a well-controlled criterion study relating these measures to various criteria of group effectiveness in a variety of leadership-group situations in industry. The independence of dimension scores has special relevance

here since each may be differentially related to such criteria, depending on the situation. Although such a criterion study was beyond the scope of the present investigation, correlations were obtained between some of the questionnaires and labor grievance rates in 23 departments over an eight-month period. In this limited study only one correlation reached the 1% level of significance (based on an N of 23 departments). This was the correlation of  $-.53$  between the mean scores of foremen in each department on the "Consideration" dimension of the form "What Your Boss Expects of You." The correlation with the "Initiating Structure" score of this form was  $.32$ . The trend was for departments with high worker grievance rates to be those whose foremen perceived their own supervisors as expecting them to lead with a low degree of "consideration" and a high degree of "structuring." These results, of course, are purely suggestive. An adequate evaluation of the value of these instruments in predicting group effectiveness must await additional research.

The Leadership Opinion Questionnaire has been found of value in the evaluation of a leadership training course for foremen and in the study of certain social factors affecting the foreman's leadership role (1, 3).

### Summary

The development of questionnaires to measure certain aspects of leadership attitudes in industry has been described. The questionnaires were designed to measure two relatively independent dimensions of leadership attitudes. These dimensions were called "Consideration" and "Initiating Structure." Various forms of the questionnaires, revised on

the basis of a pilot study, were administered at various levels in the industrial hierarchy. On each questionnaire, the dimensions were shown to have sufficient reliabilities, insignificant intercorrelations with each other, and adequate distributions.

A comparison of the leadership attitude scores at four plant levels revealed that the higher people were in the plant hierarchy, the less "Consideration" they felt the workers should get and the more "Structuring" they felt should be initiated. The attitudes of the foreman group on each dimension fell somewhere between what the workers expect and what their own supervisors expect, but were much more like the attitudes of their supervisors.

A comparison also was made between the attitudes of the supervisors of foremen toward leading foremen and toward leading workers. The results showed that these supervisors scored significantly higher in the amount of "Consideration" they felt should be shown the foremen relative to that shown to workers, but no significant differences in their "Structuring" attitudes toward each group. High correlations were found between these attitudes of supervisors toward foremen and toward workers on both dimensions.

With reference to the workers' attitudes concerning the amount of "Consideration" they would like in an "ideal foreman," the results indicate this depends to a large extent on the particular work group. There were significant differences between work groups relative to that within work groups in the amount of "Consideration" desired, but insignificant differences with respect to the amount of "structuring" desired. The workers as a whole were more homogeneous in their attitudes about this latter dimension.

Based on limited data, it was found that departments with high worker grievance rates contained foremen who perceived their own supervisors as expecting them to lead with a

lower degree of consideration and a higher degree of structuring.

It should be stressed that the findings reported here are regarded as specific to the particular plant and the groups of workers and supervisors studied. Additional research is needed before valid generalizations can be made. It is possible that future research will indicate that combinations of measures of such things as group characteristics, needs and expectations, leadership attitudes, behaviors and perceptions, pressures from supervisors, etc. can yield more successful predictions where ordinary testing procedures have failed in the complex field of leadership and group effectiveness.

Received August 4, 1952.

### References

1. Fleishman, E. A. *Leadership climate and supervisory behavior*. Columbus, Ohio: Personnel Research Board, Ohio State University, 1951.
2. Fleishman, E. A. The description of supervisory behavior. *J. appl. Psychol.*, 1953, 36, 1-6.
3. Fleishman, E. A. The leadership role of the foreman in industry. *Engineering Experiment Station News*, Ohio State University, 1952, 24.
4. Halpin, A. W., and Winer, B. J. *Studies in aircrew composition III: The leadership behavior of the airplane commander*. HRRL Contract, Technical Report No. 3, Personnel Research Board, Ohio State University, May 1952.
5. Hemphill, J. K. *Leader behavior description*. Columbus, Ohio: Personnel Research Board, Ohio State University, 1950.
6. Hemphill, J. K., and Westie, C. M. The measurement of group dimensions. *J. Psychol.*, 1950, 29, 325-342.
7. Nelson, C. W. The development and evaluation of a leadership scale for foremen. Ph.D. Thesis, 1949, University of Chicago.
8. Seeman, M. *A status factor approach to leadership*. Columbus, Ohio: Personnel Research Board, Ohio State University, 1950.
9. Shartle, C. L. Leadership and executive performance. *Personnel*, 1949, 25, 370-380.
10. Stogdill, R. M., and Shartle, C. L. Methods for determining patterns of leadership behavior in relation to organizational structure and objectives. *J. appl. Psychol.*, 1948, 32, 286-291.

## Productivity and Attitude Toward Supervisor \*

C. H. Lawshe and Bryant F. Nagle

*Occupational Research Center, Purdue University*

Of utmost importance in the study of human behavior are the factors which motivate individuals. Inquiries into the motivations of people and the relations of these motivations to performance are being initiated and expanded throughout the field of psychology. This is especially true in industrial psychology.

In the industrial situation psychologists are asking, what motivates employees toward productive effort? How do the financial rewards of work, the behavior of the supervisor, the nature of the job itself, and the goals of management affect the effort of employees (4)? The most common approaches to these questions have been through the use of questionnaires, interviews and projective techniques to determine the attitudes of employees. Psychologists are seeking to relate employee attitudes to the actual industrial practices of paying for services, of supervising, of performing the job, and of setting goals that will result in the highest productivity. Implicit in this approach is the assumption that productivity is related to employee attitude. This assumption is accepted by nearly everyone, yet little experimental evidence has been presented. This paper is a report on the relationship between employee attitudes and productivity.

### Subjects

The population used in this study is part of the office force of a large industrial plant and is divided into a number of departments. Since some of the departments in the plant are small and since some did not participate in all phases of the study, this report is based

on only 14 work groups. Of the 223 non-managerial, salaried employees in these 14 work groups, 208, or 93%, participated in the portion reported here.

### Productivity Criterion

*The Rating Procedure.* Since each of the 14 work groups was engaged in a different type of activity it was impossible to get comparable objective measures of output. Instead, a paired comparison rating of productivity was used. Six executives in the plant (1. Works Manager, 2. Training Director, 3. Staff Assistant to Works Auditor, 4. Staff Assistant to Works Manager, 5. Assistant Works Auditor, 6. Works Auditor) were asked to indicate those work groups which they felt capable of rating. The range of selections was from eight to 14. The executives were supplied with paired comparison forms and instructed to indicate "... The department in each pair which is, in your opinion, doing its job better."

Each executive's ratings were converted to standard scores as suggested by Lawshe, Kephart and McCormick (6). The standard scores for each work group as given by the odd numbered raters were averaged and correlated with the mean standard scores of the even numbered raters. The resulting coefficient of .78 was stepped up by the Spearman-Brown formula to estimate the reliability of the means of all six raters, yielding an  $r$  of .88.

*Measuring Productivity.* Previous attempts of researchers to measure the productivity of work groups have generally followed two lines.

1. One approach (4) has sought to find situations in which there are comparable work groups, groups performing the same kind of work under the same conditions. Measures of productivity for the various work groups can be directly compared, since each group is doing the same job with the same equip-

\* For the past three years the Purdue Research Foundation and Louisville Works of the International Harvester Company have been cooperating in personnel research. This report is concerned with only one phase of a larger study involving the relationships between work group productivity, employee attitudes, and supervisory sensitivity to employee attitudes. A complete report of the study will appear later in the literature.

ment. In the best known application of this approach (4) the productivity measures for the various groups had little variability, and the resulting relationships with measured attitudes were small. While this approach to group productivity has many advantages, it is limited by the rarity with which one finds a number of work groups comparable in size, work performed, physical working conditions, equipment, and financial rewards.

2. Another approach (2) has sought to determine how well each work group meets its output quota. A "normal" level of output is set for each work group, and the group's relative productivity is represented by the actual output divided by the normal level prescribed. This approach is limited not only by the validity of the output levels prescribed, but also by the fact that the productivity of many work groups, especially in an office situation, can rarely be measured in physical units turned out, either because it is not the job of the particular group to process so many of this or that, or because the work output is regulated by the activities of other departments.

*Rating Limitations.* The use of a rating approach to work group productivity as utilized in this study also has its limitations. It has the prime limitation of any rating system; one does not know for sure what the raters really had in mind when they rated. In this case an effort was made to stress the *relative* performance of the work groups. Verbal association of a supervisor with his particular work group was avoided in an effort to minimize the influence of supervisor personality in the ratings. It is the feeling of the authors that the raters thought of the various work groups as functioning entities which were there to serve the organization. How little trouble the work group caused, whether or not it had the answers when called upon, whether or not it could cope with rush situations, and similar considerations are believed to have been the prime factors in the executives' ratings.

#### Attitude Toward the Supervisor

*The Questionnaire.* Nearly all office employees of the plant filled out a tailor-made

attitude questionnaire. The questionnaire included 21 items about the individual employee's immediate supervisor so as to provide, to some extent, a diagnostic view of the supervisor as well as a total score representing the employee's opinion of the supervisor.

*Item Selection Procedure.* Using a primary group composed of 50% of the participating employees, the 21 questions were scored by giving a weight of 1 to the most favorable response and 0 to the other one, two or three responses. Total scores were computed for each employee. On the basis of total score the primary group was divided into a high-scoring half and a low-scoring half. Then the per cent of the high-scoring half giving the most favorable response to each of the 21 items was computed. The significance of the difference between these two per cents was computed for each item by means of the Lawshe-Baker nomograph (5). Two items were discarded by this process. All remaining items on the survey were also processed in the manner. Three new items were added to the 19, making a total of 22 items measuring employee attitude toward the supervisor.

*Scale Reliability.* Each questionnaire in the holdout group was scored on the 22 items. Separate total scores for each employee were computed for the 11 odd numbered items and for the 11 even numbered items, and these were correlated. The resulting coefficient of .865, when stepped up for 22 items, yielded a scale reliability of .92.

Individual employee scores on this scale ranged from 0 to 22. This score may be easily interpreted as the number of questions which the employee answered in the most favorable manner. Average scores for attitude toward each supervisor were computed and ranged from 8.8 to 19.3. The average score of the 14 work groups toward their supervisor was 13.9 and the standard deviation 3.2.

*The Attitude Dimension.* The content of the 22 items is important to an understanding of what was being measured by the scale. The questions covered many aspects of the supervisor's behavior as perceived by the employees, including such things as, does he: give you straight answers, avoid you when he

knows you want to see him about a problem, criticize you for happenings over which you have no control, delay in taking care of your complaints, keep you informed, give you recognition, show interest in your ideas, follow through on his promises, explain to you the "why" of an error to prevent recurrence, give you sufficient explanation of why a work change is necessary, etc.

### Results

**Correlation.** The average rating of each work group on how well it was doing its job was correlated with the average attitude score in the work group toward the supervisor. The Pearson coefficient was .86. With 12 degrees of freedom, a correlation of .661 is significantly different from zero at the 1% level of confidence. Figure 1 provides a visual indication of the relationship between the two variables as well as an indication of the general dispersion of each variable.

**Interpretation.** In the interpretation of this very high relationship between rated productivity and employee attitude toward supervisor, caution must be exercised. It is all too easy to fall into the error of cause and effect thinking. On the basis of this study it

can be concluded only that the behavior of the supervisor, as perceived by the employees, is highly related to the productivity of the group as perceived by higher management.

The literature has long been replete with statements as to the influence of the supervisor on group output. French says, "Leadership has long been regarded as the most important factor in group effectiveness . . ." (1, p. 475). He points out that, "Since the manipulation of (or allowance for) variables related to morale is in institutional groups primarily the responsibility of appointed leaders, the factor of leadership assumes central significance" (1, p. 485). If the basic assumption that the attitudes of people exert a great influence over their performance is true, then it follows that the leader has within his power a means by which he can influence the output of the group.

Results similar to those reported here have been obtained in the Prudential study. Katz (3) lists a number of variables in supervisory behavior which were related to the productivity of the work group. It was found that supervisors of high productivity groups placed less direct emphasis on production as the goal, encouraged worker participation in making decisions, were more employee centered, and spent more time in supervision and less in production work. The only employee attitude in the study positively related to productivity was pride in the work group. However, employee attitude toward supervisor was not mentioned in the report. In view of the types of supervisory behavior which were found to be related to productivity, it might be inferred, however, that employee attitude toward this supervisory behavior would also have been so related.

### Summary

A measure of relative productivity of work groups in doing their jobs was obtained by having six executives rate the work groups by the paired comparison system. An attitude questionnaire was administered to the employees of these work groups. From this questionnaire 22 items were used to measure employee attitude toward the immediate supervisor of the work group. The correla-

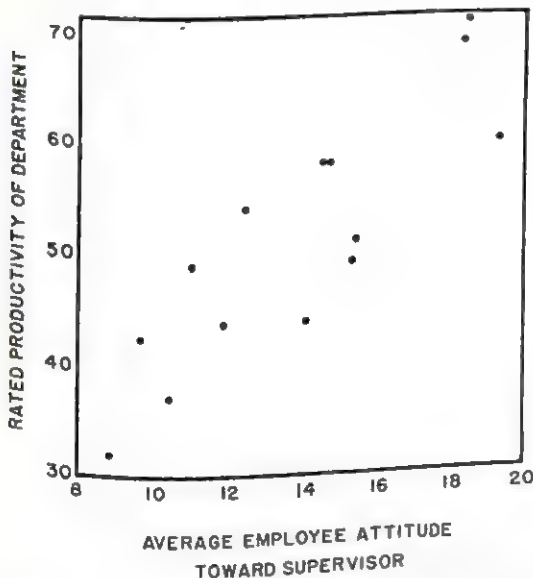


FIG. 1. Scatter diagram showing relationship between productivity of departments as rated by six executives and average employee attitude toward supervisor. The correlation between the variables is .86.

tion between the executives' rating of the productivity of the work groups and the employee's attitude toward the supervisor was .86. This relationship substantiates the hypothesis that the supervisor's behavior, as perceived by the employees, is highly related to the output of the work group.

Received April 14, 1953.

Early publication.

#### References

1. French, R. L. Morale and leadership. In National Research Council, Committee on Undersea Warfare, Panel on Psychology and Physiology, *A survey report on human factors in undersea warfare*. Washington: National Research Council, 1949. Pp. 463-490.
2. Giese, W. J., and Ruter, H. W. An objective analysis of morale. *J. appl. Psychol.*, 1949, 33, 421-427.
3. Katz, D. Morale and motivation in industry. In W. Dennis (Ed.), *Current trends in industrial psychology*. Pittsburgh, Pa.: University of Pittsburgh Press, 1949.
4. Katz, D., Maccoby, M., and Morse, Nancy C. *Productivity, supervision and morale in an office situation. Part I*. Ann Arbor, Mich.: Institute for Social Research, University of Michigan, 1950.
5. Lawshe, C. H., and Baker, P. C. Three aids in the evaluation of the significance of the difference between percentages. *Educ. psychol. Measmt.*, 1950, 10, 263-270.
6. Lawshe, C. H., Kephart, N. C., and McCormick, E. J. The paired comparison technique for rating performance of industrial employees. *J. appl. Psychol.*, 1949, 33, 69-77.

## A Simplified Procedure for the Measurement of Employee Attitudes

Melany E. Baehr

*Industrial Relations Center, University of Chicago*

During the past several years, the Industrial Relations Center of the University of Chicago has been engaged in the development of the SRA Employee Inventory, an instrument for assessing employee attitudes.<sup>1</sup> The inventory was developed by a coordinated research team representing the fields of psychology, sociology, business, and economics. It was the intent of the research team to construct an inventory which would yield a profile of scores to reflect the attitude of any given group of employees toward the significant factors in the work situation. In addition, the inventory was to be so constructed that its administration and scoring and the interpretation of results could be accomplished with the minimum expenditure of time and effort.

During the developmental stage of the inventory, the writer investigated experimentally several problems in test construction. This investigation dealt with the way in which the profile of scores was affected by: (1) the arrangement of the items (randomized vs. categorized items); (2) the number of scale intervals (five-point vs. three-point scales); and (3) the scoring procedure (unweighted vs. weighted responses).

The effect on the profile of scores was investigated separately for each of these conditions. In addition, a comparison was made of the profiles of scores resulting from six possible combinations of item arrangement, number of scale intervals, and scoring procedure. These six combinations represent procedures of an increasing degree of complexity in the treatment of the employee responses. The object was to identify the simplest procedure which could be used without loss of information.

### The Problem

*Randomized vs. Categorized Items.* In an inventory or test in which groups of items are combined to yield sub-test (category) scores, the items may be presented either in random order or grouped under the category headings to which they belong. The question arises as to whether or not the grouping of items will affect the profiles of category scores. In other words, do the test items yield different profiles of category scores when they are grouped together or categorized in the inventory than when they are randomized throughout the inventory?

*Five-Point vs. Three-Point Scale.* The number of intervals which can be used effectively in any inventory or test is a function of such conditions as the degree to which the attribute to be rated can be objectively defined, the degree of skill possessed by the raters in the use of rating scales, and their interest in making the ratings. The question arises as to whether or not the use of a five-point scale would yield a profile of category scores which was different from that obtained with the three-point scale.

*Unweighted vs. Weighted Responses.* When the five-point scale is used, the further question arises as to whether or not the profile of category scores will be affected if the responses in the extreme scale intervals are given twice the weight of the responses in the scale intervals immediately preceding them.

*The General Problem.* Six procedures for the treatment of employee responses result from the combination of the three conditions discussed above. These are as follows: (1) a three-point scale with categorized items; (2) a three-point scale with randomized items; (3) an unweighted five-point scale with categorized items; (4) an unweighted five-point scale with randomized items; (5) a weighted five-point scale with categorized

<sup>1</sup> Published by Science Research Associates, Inc., Chicago.

items; and (6) a weighted five-point scale with randomized items.

The hand scoring of an inventory utilizing a three-point scale and categorized items need involve only a count of the number of acceptable responses in groups of consecutive items. Under these conditions the profile of category scores is immediately available. It is self-evident that an inventory composed of randomized items or one in which responses to items must be made in terms of a scale having a larger number of intervals, especially if the extreme intervals are weighted, would require a proportionately greater amount of administration and scoring time. The general problem, therefore, is to determine whether or not the simplest procedure (use of the three-point scale with categorized items) yields a profile of category scores which is different from the profile yielded by the more complicated procedures.

### The Experimental Design

A total of 64 items, consisting of statements descriptive of the work situation, were selected for inclusion in a preliminary form of the SRA Employee Inventory. These were grouped under the following general categories: I. Job Demands; II. Working Conditions; III. Pay; IV. Company Benefits; V. Changes on the Job; VI. Friendliness of Fellow Employees; VII. Supervisory Effectiveness; VIII. Management and Company Policy; IX. Communication; and X. Personal Satisfaction on the Job.

Four forms of the inventory were constructed as follows:

1. Randomized items to which responses were to be made on a three-point scale.
2. Randomized items to which responses were to be made on a five-point scale.
3. Categorized items to which responses were to be made on a three-point scale.
4. Categorized items to which responses were to be made on a five-point scale.

Each of the four forms of the inventory was administered to a separate group of employees at a retail store of a large merchandizing organization in Chicago. These groups

of employees were approximately equal and were randomly selected from a total experimental population of 454 subjects.

For the two forms in which the three-point scale was used, the employee was required to indicate whether he agreed, was undecided, or disagreed with each statement (i.e., inventory item). About half the items in each category were company oriented, and half, anti-company oriented. If an employee agreed with a company-oriented item, e.g., "Management here is really interested in the welfare of employees," it was regarded as a "Favorable" response (i.e., favorably oriented toward the company). If he disagreed with such an item, it was regarded as an "Unfavorable" response. The converse held with respect to the items that were anti-company oriented.

Essentially the same procedure was followed also for the two forms in which the five-point scale was used. However, since the five-point scale provided the employee with the opportunity to indicate, if he so chose, that he strongly agreed or strongly disagreed with an item, there were two additional types of response. These were regarded as indicating a "Very Favorable" or a "Very Unfavorable" orientation toward the company.

### Results

The inventory yields a profile of ten category scores. Each category score is the per cent of favorable responses made by the group to the items in the category. It is regarded as a measure of the positive feeling held toward the company by the group. The per cent of favorable responses rather than the number of favorable responses was used because the number of items varies from category to category. The specific formulae employed in the calculation of the per cent favorable response (P.F.R.) are given below.

Three-Point Scale (Categorized or Randomized Items),—

$$P.F.R. = \frac{100F}{N \times I}$$

Unweighted Five-Point Scale (Categorized or Randomized Items),—

$$\text{P.F.R.} = \frac{100(F + VF)}{N \times I}$$

Weighted Five-Point Scale (Categorized or Randomized Items),—

$$\text{P.F.R.} = \frac{100(F + 2VF)}{2N \times I}$$

P.F.R. is the per cent favorable response,  
F is the number of "Favorable" responses,  
VF is the number of "Very Favorable" responses,

N is the number of persons in the group, and

I is the number of items in the category.

The profiles of the per cent favorable response which result from the six procedures in the present investigation are shown in Figure 1, where the following identifying symbols have been used:

- A. Unweighted five-point scale with randomized items.
- A'. Unweighted five-point scale with categorized items.
- B. Weighted five-point scale with randomized items.
- B'. Weighted five-point scale with categorized items.
- C. Three-point scale with randomized items.

C'. Three-point scale with categorized items.

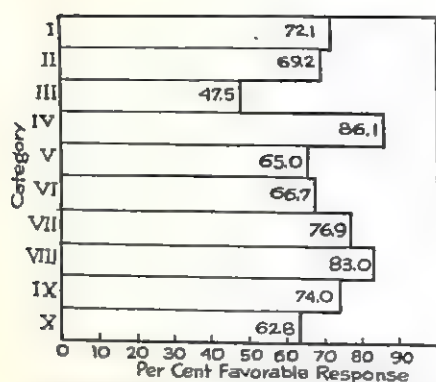
It can be seen by inspection of Figure 1 that all the profiles exhibit a great similarity in shape, though the profiles in which weighted responses were used occur at a lower level on the scale.

A quantitative measure of the similarity between the profiles was obtained by calculating the product-moment correlation coefficients between the sets of category scores. A comparison of the profiles obtained from randomized and from categorized items, when the possible effects of the number of scale intervals and the method of scoring are constant, is obtained by comparing profile A with A', B with B', and C with C'. A comparison of the profiles obtained from the three-point and the five-point scales, when the possible effects of the order of appearance of the items are constant, is obtained by comparing profile A with C, A' with C', B with C, and B' with C'. A comparison of the profiles obtained from the five-point weighted scale and the five-point unweighted scale, when the possible effects of the order of appearance of the items are constant, is obtained by comparing profile A with B, and A' with B'. For the sake of completeness, the other possible profile comparisons were also made, i.e., A with B', A' with B, A with C', A' with C, B with C', and B' with C. The results for these four sets of comparisons are given in Table 1.

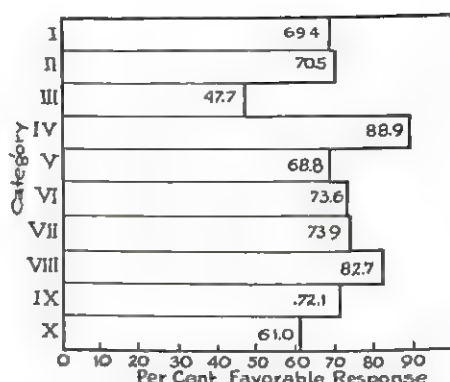
Table 1

Product-Moment Correlation Coefficients Between the Ten Category Scores in the Six Profiles  
Obtained from the Different Procedures in the Treatment of Employee Responses

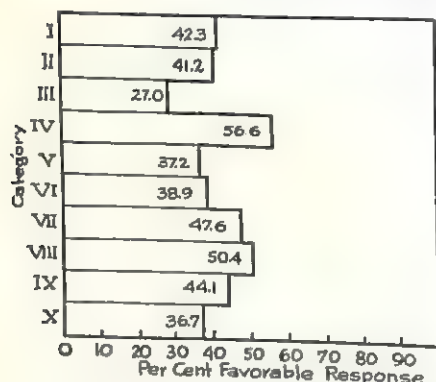
Randomized vs. Categorized Items						
Profiles Compared	AA'	BB'	CC'			
<i>r</i>	.96	.97	.97			
Five-Point vs. Three-Point Scale						
Profiles Compared	AC	A'C'	BC	B'C'		
<i>r</i>	.97	.99	.98	.98		
Unweighted vs. Weighted Responses						
Profiles Compared	AB	A'B'				
<i>r</i>	.99	.98				
Other Possible Comparisons						
Profiles Compared	AB'	A'B	AC'	A'C	BC'	B'C
<i>r</i>	.95	.95	.95	.97	.94	.98



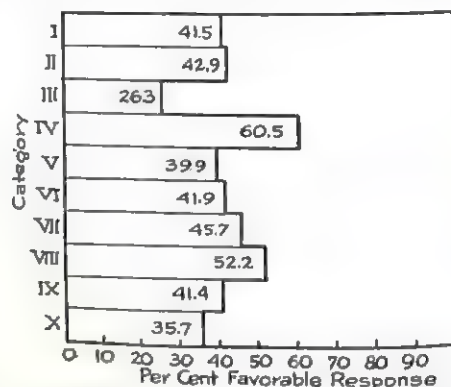
A. Five-Point Scale Randomized Items (Unweighted)



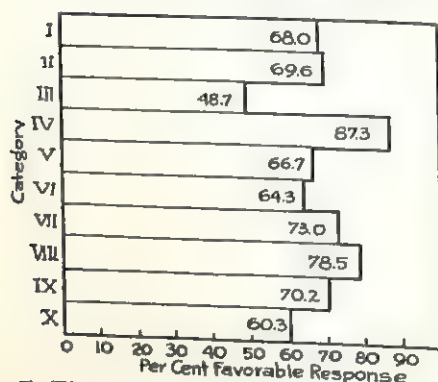
A'. Five-Point Scale Categorized Items (Unweighted)



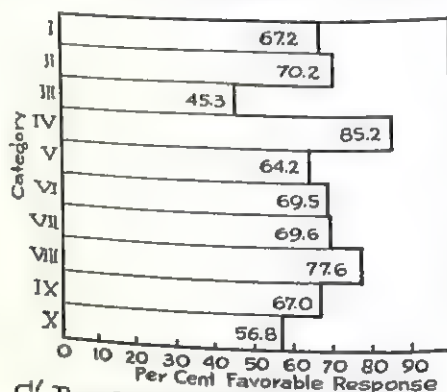
B. Five-Point Scale Randomized Items (Weighted)



B'. Five-Point Scale Categorized Items (Weighted)



C. Three-Point Scale Randomized Items



C'. Three-Point Scale Categorized Items

FIG. 1. Profiles showing the per cent favorable response to the ten categories in the Employee Inventory. There is one profile for each of the six procedures employed in the treatment of the employee responses.

The high correlation coefficients in Table 1 indicate that the profiles are similar in shape, or in other words, that there is a linear relationship between the sets of category scores contributing to the profiles. The sets of category scores contributing to the six profiles A, A', B, B', C, and C' were also compared with respect to their variances. Application of

Bartlett's test of homogeneity of variances<sup>2</sup> gave a chi-square of 1.291 with 5 degrees of freedom, yielding a *P* value of .94.

The results indicate, therefore, that the six profiles are highly similar with respect to both their shape and the variability of their cate-

<sup>2</sup> Snedecor, George W. *Statistical methods*. Ames, Iowa: The Iowa State College Press, 1946, p. 249 ff.

gory scores. The profiles obtained from the weighted five-point scale are at a different level, but can readily be converted to the same level as the remaining profiles by applying a constant stretching factor to the category scores.

The high correlations between the profiles indicate that only minor variations occur in the individual category scores. Such minor variations would not affect the interpretation of the profile as a whole.

#### Summary

A comparison was made of the profiles of category scores obtained from six different procedures in the treatment of the responses to items in an inventory designed to reflect the attitude of industrial employees toward the significant aspects of the work situation. These six procedures represented progressively increasing complexity in the arrangement of the items, the rating scales employed, and in the method of scoring the responses.

Comparison of the relevant profiles showed that:

- (1) Almost identical profiles were obtained from randomized and from categorized items,
- (2) Almost identical profiles were obtained from five-point and from three-point scales, and
- (3) Almost identical profiles were obtained from unweighted and from weighted responses.

Finally, all possible comparisons were made between the six profiles studied in this investigation. The 15 product-moment correlation coefficients which resulted ranged from + .94 to + .98 ( $N = 10$ ). This indicated that the profiles are highly similar in shape. Application of Bartlett's test of homogeneity of variances indicated that the profiles were similar with respect to the variability of their category scores. It can be concluded, therefore, that the use of the simplest procedure, i.e., the three-point scale with categorized items, results in a profile of scores which would be interpreted in exactly the same way as those resulting from the other five, more complicated procedures.

It is clear that the use of the simplified procedure will result in considerable savings in time, labor, and costs involved in the administration and scoring of the inventory. From a practical standpoint, therefore, this investigation points up the desirability of running pilot studies to determine whether or not a simpler form of a test or inventory will yield any less information than more complicated ones for specific subject populations. This is especially true when, as is often the case in industrial and educational institutions, an inventory which is once accepted is likely to be routinely administered to thousands of cases year by year.

*Received July 14, 1952.*

# The Motivation Factor in Testing Supervisors

Eugene Emerson Jennings

Wharton School of Finance and Commerce, University of Pennsylvania

Effectively using psychological testing to aid in selecting supervisory personnel presents an extremely important problem in motivation. The question is whether there are differences in motivation in taking tests for research or for actual promotion purposes. If there are motivational differences between taking tests for research and for keeps, which basis of motivation will elicit test responses that more clearly reflect the individual's actual aptitude?

## Method

The writer had an opportunity to check this with a sample of 40 supervisors who volunteered initially to participate in a testing program aimed at obtaining for research purposes a measure of the qualities and characteristics identifying the group as a whole. The supervisors were randomly divided into two groups of 20 each. Rough comparability was obtained in age, education and experience since differences between these means and sigmas did not exceed the .05 level of significance. The two groups identified as 1 and 2 were given the *Wonderlic Personnel Test Form A*.

Three months later the same two groups were given Form B, but supervisors in the Control Group 1 were encouraged to cooperate for purely research purposes while supervisors in the Experimental Group 2 were asked to cooperate for the purpose of giving management additional information for determining whom among them to promote to higher supervisory levels.

In order to determine which basis of motivation elicited test scores more nearly representative of actual aptitude, a criterion of over-all performance was obtained. Superiors knowing each supervisor in Group 1 ranked them from best to poorest on over-all performance as defined in a training session. The same procedure was followed in evaluating supervisors in Group 2. A reranking of each supervisor in both groups three months

later showed the criterion to have an estimated +.89 reliability. Correlations between test scores and criterion for Groups 1 and 2 for both testing situations were obtained by the rank-differences method.

## Results

Table 1 shows the mean scores and sigmas for Group 1 and 2 with respect to Form A and B of the *Wonderlic Personnel Test*.

Whereas the differences in means and sigmas were not significant between the first and second testing for the Control Group 1, the Experimental Group 2, believing their performance at the second testing would affect their opportunity for promotion, increased their mean score almost seven points.

Table 1  
Scores of the Wonderlic Tests

	Group 1 (N = 20)	Group 2 (N = 20)	d*
Form A			
Means	19.1	19.9	.78
Sigmas	5.5	5.0	.44
Form B			
Means	20.0	26.6	6.63†
Sigmas	5.7	6.4	.63

\* Differences computed before rounding to one decimal place.

† Indicates significant difference beyond .05 level of confidence.

However, did supervisors in both Groups 1 and 2 maintain comparable scores in the two testing situations? The correlations by the rank-differences method between first and second testings were +.76 and +.39, respectively, for Groups 1 and 2. The former but not the latter is significantly greater than zero since it exceeds the .05 level of confidence.

Generally, supervisors in Group 1 maintained comparable absolute and relative

scores in both testing situations. Supervisors in Group 2 did not maintain absolute and relative scores when advised that promotions would be based on test performance. Inspection revealed that several supervisors changed rank-positions from highest to lowest and in two cases rank values changed while numerical scores did not.

The correlations between test scores and criterion for Groups 1 and 2 were, respectively,  $+ .41$  and  $+ .34$  for the first testing and  $+ .37$  and  $+ .67$  for the second testing. Only the last correlation is significantly greater than zero since it exceeds the .05 level of confidence.

These data tend to indicate that an insignificant relationship existed between test scores and criterion of over-all performance when the tests were administered for purely research purposes. However, changing the basis of motivation from that of research to that of promotion purposes brought about a highly significant relationship between test scores and criterion.

It might be interesting to mention that two men from Group 2 were actually promoted since the several supervisors up for consideration were just by chance in Group 2. However, their test scores were not helpful in deciding which of the several to promote since all of their scores on the second test were fairly high. But had scores on the first test, given for purely research purposes, been used to aid management in promoting two supervisors, it is doubtful that the two ac-

tually selected would have been since they had two of the lowest scores in their group.

### Summary

The problem of whether there are differences in motivation in taking tests for research or for promotion purposes was studied by giving to a group of supervisors two forms of the *Wonderlic Personnel Test* with a time interval between for research purposes. A second group took the same two forms but the second administration was with reference to possible promotion. The following results were obtained:

1. The promotion motivation produced significant increases in the mean score whereas the control group showed no such increases.

2. The promotion motivation changed the individual's relative standing in the experimental group as shown by the lower correlations between the two tests than occurred in the control group.

3. Scores motivated by promotion purposes had greater validity as indicated by correlations with a criterion based on ratings of over-all performance.

Although it is very difficult to draw general conclusions, the implications of this study should serve to sound a note of caution to others doing research on aptitude tests in industry to take special pains to control the factor of motivation.

*Received February 18, 1953.*

*Early publication.*

## The Minnesota Engineering Analogies Test \*

Marvin D. Dunnette

*Industrial Relations Center, University of Minnesota*

Engineers and technically trained personnel are key figures in meeting unprecedented demands of our armed forces, defense industry, and our civilian economy. As a result, the country faces a critical shortage of engineering personnel. In the year 1949-1950, a total of 57,159 (20) engineers were graduated from the nation's technical schools. These graduates were rapidly absorbed by industry. Against an estimated annual requirement for 30,000 new engineers, the yearly crop of engineering graduates, however, is rapidly declining. Thus from a total of 38,000 graduated in 1951, the estimated number of graduates falls to 17,000 for 1954 (19).

In view of these figures, the selection of engineering students to continue in pursuit of graduate degrees becomes a problem of primary importance. It is undesirable that technical manpower be wasted in the unsuccessful pursuit of advanced training. In like manner, it is important that educators be able to identify the most able students in order that they may be urged to pursue graduate work. It has long been recognized by engineering faculties that wise selection of advanced students would be facilitated by development of a short, easily administered test with demonstrated validity for the assessment of potentialities necessary to success in graduate school. This article presents the rationale for the use of a special analogies test in this assessment task and constitutes a description of an exploratory attempt to build such an instrument.

Only within recent years have systematic efforts been made to predict success in engineering curricula. Usually such investigations have not met with the degree of success

enjoyed by projects designed to develop predictive devices in other fields of academic endeavor. However, studies (1, 3, 4, 9, 11, 16, 17) concerned with the prediction of success in undergraduate engineering have uniformly shown certain measures to be of maximum utility. It would appear that an ideal combinational measure for the evaluation of a person's aptitude for engineering would include measures of previous academic achievement (1, 4), general intelligence (16), and facility in mathematics (1, 3, 4, 9, 11, 17).

The problem with which this investigation was most concerned (i.e., the evaluation of graduating engineers) has been little recognized in the literature. The Graduate Record Examination has been used extensively in several fields, but little information relating performance on the G.R.E. to achievement in advanced engineering training is available. Learned (5, 6) in discussing the relative merits of the G.R.E. states that the theoretical approach of the G.R.E. (i.e., the testing of information gleaned from a variety of subject matter fields) has proved successful in prediction at the higher levels of academic endeavor. Speer (13), on the other hand, feels that the broad generality of the subject matter tested by the G.R.E. is the very factor which makes it unsuitable for use with engineers. He emphasizes that selection of capable engineering graduate students must include a measure of general mental ability as well as measures of achievement in previous work.

In general, little of a definitive nature has been done in the evaluation of graduating engineers. There is an indication that success in postgraduate employment is related to undergraduate grades (10). It is felt that proficiency in graduate school can be predicted best by a test combining a measure of general intelligence with some measure of previous achievement (13). One relevant study (15) has shown that tests requiring the ability to perform abstract reasoning are

\* This study was completed while the author was Teaching Assistant in the Institute of Technology, University of Minnesota, 1950-51. It was completed in partial fulfillment of M.A. requirements in personnel psychology. The author wishes to acknowledge the assistance and guidance given him by his advisor, Professor Donald G. Paterson.

efficient predictors of success in advanced study in the physical sciences.

Experience with the verbal analogy as a test item has shown it to have characteristics which would appear to make it an efficient device for use in evaluating engineering graduates. The verbal analogy item is short. A test including many such items may thus be easily administered in a brief period of time. Because the analogy requires the perception of relations and the generation of correlate relations, it is a measure of abstract intellect (12). Furthermore, although it is related to verbal facility, it has also been shown to be associated ( $r = .67, .68$ ) with measures of arithmetic reasoning and arithmetic computation (18). Factor analyses of verbal analogies tests (14) have indicated high loadings in V (verbal) and D (deductive) factors. The latter factor is most prevalent in tests calling for arithmetic reasoning, and number series completion, abilities which are important in predicting success in engineering training.

In the construction of analogy items, it is possible to include concepts calling for knowledge in specific subject matter fields. Thus analogies tests may be used to measure previous achievement. Levine's study (7, 8) bears particularly on this point. He developed an analogies test specific to the subject matter of psychology. He found his test to be a slightly better predictor of achievement in psychology courses than a test of general ability such as the Miller Analogies. He concludes, "At any rate the data obtained in this project would tend to indicate the feasibility of exploring the possible uses of special analogies tests in other fields" (8, p. 305).

Thus, a special analogies test involving engineering knowledge and concepts may be an efficient instrument in measuring capabilities necessary to success in graduate engineering.<sup>1</sup> Because of this, such a test was constructed and used in this exploratory evaluation of engineering graduates.

<sup>1</sup> It may turn out that this type of test would be even more important in the selection and placement of engineers (sales, research, design, electrical, etc.) in business and industry. For this reason, the test here reported has been extended and is now being validated in a number of established engineering research departments.

## The Present Study

The purpose of this study was to build a test applicable to all fields of engineering. This decision necessitated drawing items from that store of information which can accurately be said to comprise the "common-core" of academic knowledge among graduating engineers. An analysis of the curricula in 14 engineering colleges indicated the so-called "common-core" to consist of courses in inorganic chemistry, analytic geometry, trigonometry, algebra, differential and integral calculus, physics, hydraulics, statics and dynamics, strength of materials, thermodynamics, and a survey of the basic principles of electrical engineering. Items were written for each of these subject matter fields. Minute details were avoided; only important principles basic to the fields were included. By so doing, it was hoped that esoteric informational content would be ruled out as a determiner of item difficulty. From the initial pool of 135 items, 90 were selected for preliminary administration. The following are examples of the analogies<sup>2</sup> which were used:

Consider a triode:

Spectators:turnstile::plate current:

- (1.) cathode; (2.) plate; (3.) anode;  
(4.) grid.

Pauper:money::riveted butt joint:

- (1.) bearing stress; (2.) bending stress;  
(3.) tensile stress; (4.) shearing stress.

Diameter:circumference:: $y = bx$ :

- (1.)  $x^2 + y^2 = r^2$ ; (2.)  $x^2 = 2py$ ;  
(3.)  $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ ; (4.)  $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ .

The 90-item test was administered to 203 engineering seniors enrolled in G.E. 103, a survey course of engineering ethics, which is required of all graduating seniors in the Institute of Technology at the University of Minnesota.

Of the 203 seniors who took the preliminary form of Minnesota Engineering Analogies Test, only 91 completed every item on

<sup>2</sup> The correct response for the first example is (4.) grid; for the second is (2.) bending stress; and for the third is (1.)  $x^2 + y^2 = r^2$ .

the test. This was because of the limited time afforded (50 minutes) by the length of the class period. Each student's score was, therefore, expressed in terms of an accuracy index derived by dividing the number of items answered correctly by the number answered incorrectly. A scatter diagram portraying the relation between the number of items attempted and the accuracy index indicated that slow workers worked just as accurately as more rapid workers. Thus it was indicated that ability to finish the test was not related to a student's proficiency in the test. Therefore, it did not seem desirable to include the speed factor in the analysis of results. For item analysis purposes then, the scores were expressed in terms of the accuracy index.

Davis' item analysis chart (2) was used to compute biserial validity coefficients. Two indexes were computed for each item. The first validity index was based on an "internal" criterion, the accuracy index. The second validity index was based on an "external" criterion consisting of the over-all honor point average<sup>3</sup> earned at the University of Minnesota. A total of 63 items exhibiting validity coefficients above .10 on both criteria were combined into a final form of the test. This final form was then administered to 53 graduate students in engineering.

### Results

Among the 203 seniors, the correlation between honor point average and the accuracy index for the 90-item test was .57. A correlation of this magnitude was considered encouraging in view of the fact the test still contained many poorly discriminating items.

Table 1 shows distributions of the validity coefficients obtained in the item analysis.

The corrected odd-even reliability of the 63-item test administered to the graduate students was .86. A reliability of this magnitude compares favorably with the reliability coefficients of some of the more widely used standardized tests. The distribution of responses made by the graduate group showed

<sup>3</sup> The honor point average is calculated on the basis of three honor points for each credit of A, two for B, one for C and zero for either D or F.

Table 1  
Magnitudes of Validity Indexes Obtained

Biserial <i>r</i>	Internal Consistency Criterion (R/W)	External Validity Criterion (HPA)
	Number of Items	Number of Items
<0	4	12
0-.10	12	14
.11-.20	15	26
.21-.35	28	32
>.35	31	6
Total	90	90

that many of the distractors, apparently adequate within the senior group, failed to function effectively for graduate students. However, for the graduate group, the average Davis Difficulty Index (2) was 57. This value corresponds to a proportion of successes of .63. This indicates that in spite of the shrinkage of distractor effectiveness, the test was moderately difficult for these high ability students.

For purposes of comparison, the tests of the seniors who finished the test were re-scored for the 63 items of the final form. Figure 1 shows the distribution of scores for the two groups (seniors and graduate students) on this form. The graduate students scored markedly higher having a mean of 37.1 and S.D. of 7.24 compared with the senior mean of 28.7 and S.D. of 7.18. The critical ratio was 6.76.

In terms of overlap, only 13 per cent of the seniors exceeded the median of graduate students, and only 9 per cent of the graduate students fell below the median of the seniors. The low amount of overlap is a definite indication that the test operates in a valid manner to identify the more able engineering students.

But to what extent does the test differentiate among graduate students with different abilities? In order to investigate this question, the graduate student group was divided into first, second, and third year students according to the following plan: 1st year—0-3 quarters; 2nd year—3-6 quarters; and 3rd year—more than 6 quarters.

Table 2

Differential Performance of Seniors and Graduate Students on the 63-Item Test

Group	N	Mean	Standard Deviation	t	Probability Level
Seniors	91	28.7	7.18	3.88	P < .001
1st Year Grad. Students	24	34.6	6.26	.32	P = .75
2nd Year Grad. Students	13	35.3	5.63	2.84	P < .01
3rd Year Grad. Students	16	42.4	7.05		

Figure 2 shows the distribution of scores within each of these three groups. The first and second year students showed similar performance, but the third year students exhibited marked superiority. These results are important when it is remembered that third year students include only the carefully screened Ph.D. candidates; the other two groups consisting of master's candidates. Table 2 summarizes the performance on the 63-item test of all four groups (seniors, 1st, 2nd, 3rd year graduate students). In terms of overlap, only 13 per cent of the students in the first two years of graduate school reached the first two years of graduate school students. or exceeded the median of 3rd year students. In like manner, only 19 per cent of the latter

group fell below the median of the former. These results provide further impressive evidence of the validity of the 63-item test. To the extent that candidates for the Ph.D. degree are, as a group, more able than other students in graduate engineering, the ability of the test to identify the more competent group is established.

It may be concluded that the exploratory use of the special analogy test has proved it to be a feasible device for the evaluation of engineering graduates. This conclusion is based on the fact that a large number of the analogy items discriminated sharply between academically superior and inferior students. Further support is drawn from the finding of

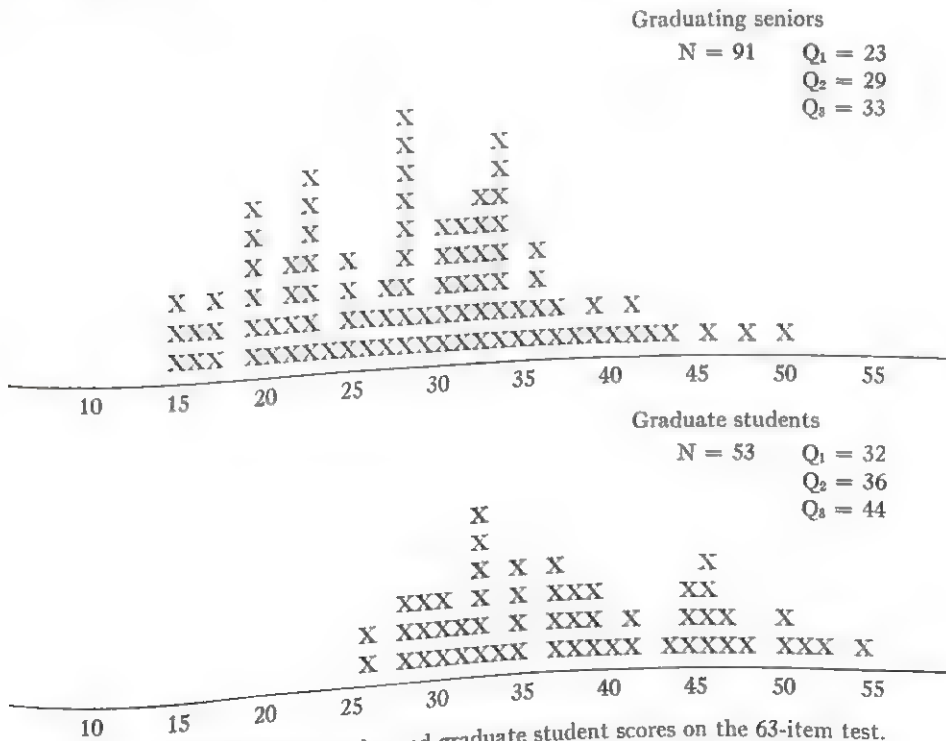


FIG. 1. Distribution of senior and graduate student scores on the 63-item test.

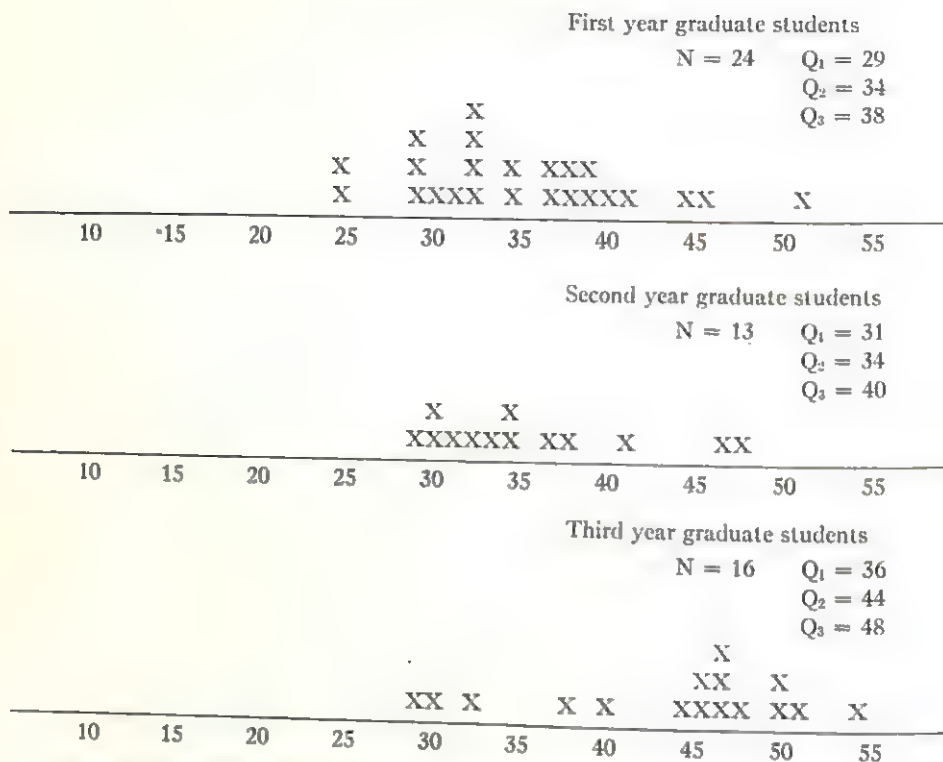


FIG. 2. Differential distribution of graduate student scores on the 63-item test.

significant differences between graduating seniors and graduate students and between graduate student groups with varying levels of ability. It is suggested that further investigation of the special analogy test may be most profitable in the development of instruments for the assessment of high level engineering abilities.

### Summary

Studies have shown that degree of success in engineering studies can be most effectively predicted by a combination of measures including previous academic performance, general intelligence, and mathematical facility. The nature of the verbal analogy item suggested that it is peculiarly fitted to the task of assessing the above attributes.

With this in mind, a 90-item engineering analogies test was constructed and administered to 203 engineering seniors. Item analyses were made using internal consistency and external validity criteria. The 63 most discriminating items were combined and administered to 53 graduate students. The odd-

even reliability of this form was found to be .86 for these highly selected graduate students.

The results for the 63-item test were examined with respect to comparisons within and between graduate students and seniors. The performance of graduate students was markedly superior to that of the seniors. Within the graduate group, the performance of third year students (Ph.D. candidates) was superior to that of first and second year students (M.A. candidates).

It was concluded that a special analogies test effectively assesses engineering abilities.

Received June 30, 1952.

### References

1. Berdie, R. F., and Sutter, N. A. Predicting success of engineering students. *J. educ. Psychol.*, 1950, **41**, 184-190.
2. Davis, F. B. *Item analysis data*. Cambridge, Mass.: Graduate School of Education, Harvard, 1949.
3. Griffin, C. H., and Borow, H. An engineering and physical science aptitude test. *J. appl. Psychol.*, 1944, **28**, 376-387.

4. Jones, V. Prediction of student success in an engineering college. *Amer. Psychologist*, 1948, 3, 295.
5. Learned, W. S. *Measurement of student knowledge as a basis for graduate study*. Carnegie Foundation for the Advancement of Teaching, Thirty-third Annual Report, pp. 62-68.
6. Learned, W. S. *The Graduate Record Examination; a memorandum on the general character and use of the examination including a summary of initial studies of its validity*. Carnegie Foundation for the Advancement of Teaching.
7. Levine, A. S. *A psycho-analogies test as an evaluation instrument for psychology students*. Ph.D. dissertation, University of Minnesota, 1950.
8. Levine, A. S. Minnesota Psycho-analogies Test. *J. appl. Psychol.*, 1950, 34, 300-305.
9. Lord, F. M., Cowles, J. T., and Cynamon, M. The Pre-engineering Inventory as a predictor of success in engineering colleges. *J. appl. Psychol.*, 1950, 34, 30-39.
10. Pierson, G. A., Jr. School marks and success in engineering. *Educ. & Psychol. Measmt.*, 1947, 7, 612-617.
11. Sackett, R. L. Discovery of engineering talent. *J. Engng. Educ.*, 1944, 35, 180-183.
12. Spearman, C. *The nature of intelligence and the principles of cognition*. London: McMillan and Co., Ltd., 1927.
13. Speer, G. S. The use of the Graduate Record Examination in the selection of graduate engineering students. *J. Engng. Educ.*, 1946, 37, 313-318.
14. Thurstone, L. L. Experimental study of simple structure. *Psychometrika*, 1940, 5, 153-168.
15. Travers, R. M. W., and Wallace, W. L. The assessment of the academic aptitude of the graduate student. *Educ. & Psychol. Measmt.*, 1950, 10, 371-379.
16. Treumann, M. J., and Sullivan, B. A. Use of the Engineering and Physical Science Aptitude Test as a predictor of academic achievement of freshmen engineering students. *J. educ. Res.*, 1949, 43, 129-133.
17. Vaughn, K. W. The Yale Scholastic Aptitude Tests as predictors of success in the college of engineering. *J. Engng. Educ.*, 1944, 34, 572-582.
18. Weisenburg, T., Roe, A., and McBride, K. E. *Adult intelligence*. New York: Commonwealth Fund, 1936.
19. Wood, Helen, and Cain, R. W. *Effect of defense program on employment outlook in engineering*. Supplement to bulletin No. 968. United States Dept. of Labor, August, 1951.
20. *World Almanac & Book of Facts for 1952*. New York World Telegram & Sun. Pp. 580.

## The Humm-Wadsworth Temperament Scale as an Indicator of the "Problem" Employee

A. R. Gilliland and S. E. Newman \*

*Northwestern University*

The Humm-Wadsworth Temperament Scale was first published in 1935. Its primary purpose was for use as an aid in industrial selection. The scale consists of 318 questions, but of these only about half are scored. The others are for use as a "setting" for the scored items. The scale is based upon the Rosanoff classification of personality and purports to measure seven different components. The test was standardized on seven groups of subjects each representing a relatively pure type of that component. A highly complicated method of scoring and validation has been devised. Suffice it to say that the split half reliability of the various components varied from .70 to .90 and the validity as checked against new criterion groups was .85 to .98 as reported by the authors (1). The reliabilities have been rechecked by Dysinger (4) by the test-retest method and found to be even higher than those reported by the authors. Most of the components are independently variable; only two components, manic and depressive, show high intercorrelations ( $r = .88$ ).

The scale has been widely used with college students, psychotic groups, and in industry. No attempt will be made to review all these studies. Reed and Wittman (5) gave the scale to 477 Elgin Hospital patients and compared the scores with a normal control group. Only the normal and cycloid components were significantly different for the two groups. Dorcus (3) used the scale with an industrial group. He reports that it correctly diagnosed 73% of the poor group and 65% of the superior group.

In the present study the Humm-Wadsworth scale was given to the employees of a relatively large industrial organization employing largely "white collar" workers. The scale was administered approximately ten years ago and the evalua-

tion was made about nine years later. The scales for 405 employees who constituted those with surnames beginning with letters before "I" in the alphabet were scored. This should constitute a random sample from approximately half the population. Of this group, 191 were still employed and rated as "successful" or "satisfactory." Another group of 139 had withdrawn from the company but without any unfavorable service record. Another group of 75 had either been dismissed or resigned while on probation. These are classified as "undesirable." Using a method of score evaluation as nearly as possible like that used by Humm (2) in his study of Los Angeles policemen, and checking the method further, as best we could in a personal conference with the author, the 405 employees were classified in terms of their Integration Index and Component Control Measure into five groups—very good risks, good, questionable, poor, and very poor. Those with Integration Indices and Component Control Measures all above 5, for example, were classified as very good risks. Those with at least two ratings as low as 1 were called very poor risks.

Of the employed group of 191 still with the company and doing satisfactory work, 9.4% received a very good rating on the test and 5.7% received a very poor scale rating. Of the 139 no longer employed but with no evidence about their success, 12.2% were rated very good by the test and 5.8% as very poor. Of the 75 who had been dismissed or withdrew for cause, 12.0% were classified as very good risks by the scale and 5.4% as very poor risks. Thus it is apparent that these results show no difference between the three employed groups in terms of scores and the scale.

As another method of evaluation, the data were arranged in a  $3 \times 5$  table with three groups of employees in terms of their work record as one axis and five degrees of success on the scale in terms of the Integration Index and Component Control Measure as the other axis. From this table a chi square as a test of deviation from the null hypothesis was calculated. A chi square of 5.93 was obtained.

\* Human Resources Research Center, Keesler AFB, Miss.

With eight degrees of freedom these data gave a  $p$  of .65. That is, so great a difference as this would occur by chance 65 times out of 100. When the middle group who had left the company between the time of testing and the time the study was made was omitted, no important change in the relationship was apparent.

Inspection of the seven components of the test showed no component in which there was a significant difference between the satisfactory and unsatisfactory employees. An examination of the Integration Index which is a summary value obtained from the test also showed no difference between these two groups.

The failure of the test to differentiate the satisfactory from the unsatisfactory workers by any of the above methods may be due to any one or a combination of the following: (1) the test may not adequately measure the components it purports to measure; (2) these components may not be essential elements for success in this industry; and (3) the company cannot distinguish between satisfactory and unsatisfactory workers.

This study does not prove that the Humm-Wadsworth scale may not be successful in selecting workers in some industries but it certainly gave no evidence of success in this situation.

Received August 4, 1952.

#### References

1. Humm, D. G., and Humm, Katherine A. Validity of the Humm-Wadsworth Temperament Scale with consideration of the effects of the subjects' response bias. *J. Psychol.*, 1944, 18, 56-65.
2. Humm, D. G., and Humm, Katherine A. Humm-Wadsworth Temperament Scale appraisals compared with criteria of job success in the Los Angeles Police Department. *J. Psychol.*, 1950, 30, 63-75.
3. Dorcus, R. A brief study of the Humm-Wadsworth Temperament Scale and the Guilford-Martin Personality Inventory in an industrial situation. *J. appl. Psychol.*, 1944, 28, 302-307.
4. Dysinger, D. W. A critique of the Humm-Wadsworth Temperament Scale. *J. abn. soc. Psychol.*, 1939, 34, 73-83.
5. Reed, P. H., and Wittman, P. "Blind" diagnoses on several personality questionnaires checked with each other and the psychiatric diagnoses. *Psychol. Bull.*, 1942, 39, 592.

## The Prediction of Success and Failure in Elementary Foreign Language Courses

Harold C. Peters \*

*The Pennsylvania State College*

This study is concerned with the prediction of success and failure in the elementary courses in French, Spanish, and German at the Pennsylvania State College. Predictions were made on the basis of scores on the Pennsylvania State College Academic Aptitude Examination (3), parts one and two. Separate predictions were made for each of the above mentioned languages.

### Procedure

*The Subjects.* The subjects in this study were all the freshmen in the Pennsylvania State College who were enrolled in the elementary courses in French, Spanish, and German in September 1951, and had taken the Pennsylvania State College Academic Aptitude Examination. The total number of subjects is 443 divided among the three languages in the following manner: (1) French—47; (2) Spanish—189; and (3) German—207.

Since the study is directed toward predicting success it was felt that freshmen would be the best subjects. According to Feder (2) the function of prediction in education is to facilitate guidance, and, if it can be effective, educational guidance is most valuable when applied to freshmen who are beginning their academic careers.

*The Criterion.* The criterion used for success was the teachers' grades in the three foreign language courses. The grades at the Pennsylvania State College range from - 2 to 3 with the 3 being the highest possible grade which can be attained in a course, and the - 2 being the lowest failing grade. A grade of - 1 is also given and this too is a failing grade. Since it was felt that those students who received a grade of 0 (the lowest passing grade) had not achieved more

than the barest minimum of success, the students who received such a grade were placed in the failing group. The composition of the groups then becomes as follows: (1) Passing—grades 3, 2, and 1; and (2) Failing—grades - 2, - 1, and 0.

*The Predictive Instrument.* The instrument used in this study was the Pennsylvania State College Academic Aptitude Examination, parts one and two. The verbal nature of these two parts (vocabulary and paragraph reading) does not necessarily indicate that they would be particularly useful in the prediction of success in the study of a foreign language. However, it seems that the two or more skills or achievements involved in these tests might be expected to have a direct relationship to language skills. All items on both tests are of a multiple choice nature with five possible choices.

Bernard (1) feels that the "learning of a foreign language consists fundamentally in the acquisition of an additional set of symbols for old, familiar meanings. . . . Since the most pressing need for the student is the knowledge of the meaning of these new symbols, the preponderant importance of vocabulary becomes at once apparent." Symonds (4) states that "presumably there should be some relationship between the size of English vocabulary and the ability to learn a new language."

Considering the importance of vocabulary in the learning of a foreign language it is felt that the measurement of the skill involved in learning a vocabulary will be of some value in predicting success in learning a foreign language. Since this cannot be done directly, we used a measure of proficiency in English vocabulary as being indicative of the ability to learn vocabulary, with the feeling that the same skills involved in developing the English vocabulary are operating in learning the vocabulary of a foreign language.

\* The author wishes to express his sincere appreciation to Dr. William U. Snyder and Dr. Ila H. Gehman under whose direction this study was done.

The paragraph reading part of the test indicates not only the subject's ability to read a paragraph, but also his ability to understand what he has read, as measured by his answers to a set of questions concerning the subject matter of the paragraph. In being able to understand the meaning of a paragraph the subject must have an adequate vocabulary, must be able to learn the meanings of new words from their context, and must have some knowledge of grammar. These skills are of importance also in the learning and mastery of a foreign language.

*Experimental Design.* The procedure used in this study was as follows: The grades of all freshmen registered in the courses in elementary French, Spanish, and German at the Pennsylvania State College were collected from the respective language departments. The grades were divided into two groups, with an equal number of grades of each language in both groups. This was done by selecting every other student (in each language group) from a list of students, arranged in order of descending magnitude of grades, and placing him in one group. The other group was composed of the remaining students. This method insured an approximately equal distribution of grades for each group. Of the 507 students in these two groups, 443 had taken the Academic Aptitude Examination and their scores had been recorded by the psychology department. These scores were then collected and the analysis begun.

What was sought was a point, or score, on the Academic Aptitude Examination (parts one and two) below which are found those students who cannot make passing grades (a grade of 0 was considered to be a failure for reasons previously discussed) in their language courses.

This cut-off point was determined on one of the two groups (many points were tried and the one yielding the best prediction was selected) and its validity was determined by testing its applicability to the students comprising the second group. Different cut-off points were located for each language, with the expectation that they would be approximately the same for all the languages.

Each of the two parts of the test was used separately in finding the cut-off points, and the two tests were also combined and a cut-off point on the combined score was located for each language.

*Statistical Analysis.* The statistic which was computed was the significance of difference between per cents failing above and below the various cut-off points which were established (*t* test).

### Results

The vocabulary test proved to be a valid differentiator between students who achieved success, and students who failed in foreign language study. Table 1 gives the results of the application of the cut-off scores to the test group in each of the three languages. It can be seen that the vocabulary test was most successful in distinguishing between successful and unsuccessful students when applied to the Spanish group. Here, of the students who fell below the cut-off score 76.9% failed, while only 35.6% of those exceeding the cut-off score failed the Spanish course. This difference was significant at the .001 level. Only with French is the difference not significant at the one per cent level.

To determine the validity of the vocabulary test as a predictor, the various cut-off scores were applied to the second group (the cross validation group). Table 1 shows that there was no decrease in the significance of difference in the Spanish group, but in the other languages a decrease was found.

The lack of significance in the French group can probably be attributed to the comparatively small number of subjects in the test and validation groups of this language. The effect of these small numbers is revealed in the relatively high  $\sigma$ 's of difference found in these groups. These were found to be almost double those of the other groups. This weakness was found to be operating when predictions were made with either of the tests as well as when the tests were combined to form a single predictor.

The paragraph reading test proved to be most efficient in predicting success and failure in German. Of the students who fell below the cut-off score, 63.8% failed German.

Table 1  
The Predictors of Success and Failure

	Vocabulary						Paragraph Reading						Combined Tests					
	Test Group			Val. Group			Test Group			Val. Group			Test Group			Val. Group		
	Fr.	Sp.	Ger.	Fr.	Sp.	Ger.	Fr.	Sp.	Ger.	Fr.	Sp.	Ger.	Fr.	Sp.	Ger.	Fr.	Sp.	Ger.
Score	60	54	62	60	54	62	29	27	29	29	27	29	92	79	87	92	79	87
N	24	97	104	23	92	103	24	97	104	23	92	103	24	97	104	23	92	103
N above score																		
Failing	2	16	15	3	9	16	2	14	11	3	10	16	1	15	16	2	8	17
Passing	9	29	32	7	27	33	7	22	35	7	26	32	7	27	35	7	28	34
N below score																		
Failing	8	40	33	6	41	31	8	42	37	6	40	31	9	41	32	7	42	30
Passing	5	12	24	7	15	23	7	19	21	7	16	24	7	14	21	7	14	22
% failing																		
Above score	18.1	35.6	31.9	30.0	25.0	32.7	22.2	38.9	23.9	30.0	27.8	33.3	12.5	33.8	31.4	22.2	22.2	33.3
Below score	61.5	76.9	57.9	46.2	73.2	57.4	53.3	68.9	63.8	46.2	71.4	56.4	56.2	74.5	60.4	50.0	75.0	57.7
Difference	43.4	41.3	26.0	16.2	48.2	24.7	31.1	30.0	39.9	16.2	43.6	23.1	43.7	40.7	29.0	27.8	52.8	24.4
$\sigma$ difference	17.80	9.19	9.43	17.33	9.33	9.49	16.62	10.60	8.91	17.33	9.61	9.54	17.05	9.39	9.35	17.30	9.03	9.47
				above						above						above		
P	.05	.001	.01	.1	.001	.02	.1	.01	.001	.1	.001	.02	.02	.001	.01	.1	.001	.02

Among the students who exceeded the cut-off score only 23.9% failed. This difference was found to be significant at the .001 level.

In the cross validation group, however, the greatest success in prediction was made with the Spanish group, and once again the least success was found with the French group.

Combining the vocabulary and paragraph reading tests does not yield a marked improvement in prediction of success and failure, as will be revealed by an examination of Table 1. Especially in the cross validation group are the results found to be almost exactly those attained when the two tests were used separately.

### Summary

The object of this study was to determine the efficiency of parts one (vocabulary) and two (paragraph reading) of the Pennsylvania State College Academic Aptitude Examination in predicting success and failure in the elementary courses in the modern foreign languages.

The results of this study show:

1. The greatest success in prediction was achieved with the Spanish group.
2. Success in French was most difficult to predict. This was attributed to the small number of subjects in this group.
3. There was very little difference in the efficiency of the predictive instruments. The combined tests were generally most successful and the vocabulary test probably somewhat more effective than the paragraph reading.

This study demonstrated that it is possible to predict success and failure in the modern

foreign languages. It was further demonstrated that tests of vocabulary and paragraph reading can be used to make this prediction.

It is suggested that the college administration take the responsibility for selecting a method of giving the students with low language aptitude (as measured by any of the instruments used in this study) an opportunity to derive some value out of foreign language study. This would no doubt involve some special treatment such as can be provided by special classes.

The highly significant results found in this study indicate that procedures such as these could be applied in schools other than the Pennsylvania State College. It is likely, however, that each school would find it expeditious to determine for itself, the most efficient predictive score. If necessary, individual schools could substitute other tests of a similar nature with which comparable results might be obtained.

Received July 15, 1952.

### References

1. Bernard, W. Psychological principles of language learning and the bilingual reading method. *Mod. Lang. J.*, 1951, 35, 87-96.
2. Feder, D. D. An evaluation of some problems in the prediction of achievement at the college level. *J. educ. Psychol.*, 1935, 26, 597-603.
3. Moore, B. V., and Castore, G. F. *The Pennsylvania State College Academic Aptitude Examination*, 1947 revision. The Penn. State Coll., 1948.
4. Symonds, P. M. *A modern foreign language test*. Publications of the American and Canadian Committees on Modern Languages, vol. 14. New York: The Macmillan Co., 1920.

## Predicting Grades in Advanced College Mathematics

John R. Kinzer and Lydia Greene Kinzer

*The Ohio State University*

This is a study of 1,244 students, of whom 78 were women, who took college algebra and of their success in subsequent courses in mathematics. The study followed these students until 29 remained at the end of seven courses in the prescribed sequence. Twenty-five of these remaining 29 students had been graduated at the time of this writing.

The seven courses in mathematics making up the sequence are briefly described as follows:

421. College Algebra. Five credit hours. . . .  
422. Trigonometry. Five credit hours. . . .  
Prerequisite, Mathematics 421. . . .

423. Analytic Geometry. Five credit hours.  
. . . Prerequisite, Mathematics 422. . . .

441. Calculus. Five credit hours. . . . Pre-  
requisite, Mathematics 423. Differentiation of  
algebraic forms, with applications; successive dif-  
ferentiation; differentiation of transcendental  
functions; parametric equations, differentials;  
curvature; theorem of mean value; indeterminate  
forms.

442. Calculus. Five credit hours. . . . Pre-  
requisite, Mathematics 441. Integration of stand-  
ard elementary forms, and integration by various  
devices; definite integrals; application to geome-  
try and physics.

443. Calculus. Five credit hours. . . . Pre-  
requisite, Mathematics 442. Numerical series  
and power series; differential equations; hyper-  
bolic functions; partial differentiation; multiple  
integrals, and applications.

601. Advanced Calculus. Five credit hours.  
. . . Prerequisite, Mathematics 443. The theory  
of limits, functions, continuity; definition and  
meaning of ordinary and partial derivatives;  
definition of definite integrals, proper and im-  
proper; fundamental theorem of the integral cal-  
culus; functions defined as integrals containing a  
parameter; mean value theorems; convergence of  
series; power series; implicit functions.

The course in advanced calculus was chosen as the culminating point because it is thought to represent the type of thinking believed to be important in the graduate study of mathematics.

The sample is 1,244 students who were enrolled in college algebra at the Ohio State University in the autumn quarter of 1946.

The data were collected after most of the students were presumed to have had time to complete the entire sequence.

In addition to course grades, percentile scores on the Ohio State Psychological Examination (OSPE) were included in the computations.

For the 29 students who completed the entire seven-course sequence some personal data are given to describe this sample in greater detail.

Table 1 presents the intercorrelation matrices of OSPE and mathematics course grades. These are presented in a manner such that one can easily see how the coefficients of correlation change as the sample decreases in size.

The means presented in Table 1 show that the better students in the early courses tend to go on into the more advanced courses. The mean OSPE gradually increases, until the last group is reached where there is a sharp upward trend. The mean OSPE of the 29 students in the advanced calculus course is 79.1 percentile.

In Table 2 are presented the regression coefficients and coefficients of multiple correlation. Although some of the coefficients in the regression equation are negative, none of the negative coefficients is significantly different from zero.

The 29 personnel cards filled out during Freshman Week by the 28 men and one woman who took Advanced Calculus were examined in an attempt to discover some clues as to success in advanced mathematics. Of the 29 persons who took seven courses in mathematics, 12 made A or B grades in Advanced Calculus, 17 made C, D or E grades. Chi-square was used to test for independence of the grade classification and the following classifications:

1. Like mathematics—like some other subject ( $\chi^2 = 0.000$ ).

Table 1

Coefficients of Correlation, Means, and Standard Deviations

Note: OSPE scores are percentiles. Grades are expressed on the basis of A = 4, B = 3, C = 2, D = 1.

	N	OSPE $\bar{x}_1$	Mathematics Courses						
			421	422	423	441	442	443	601
			$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$	$\bar{x}_6$	$\bar{x}_7$	y
OSPE	1,244 $\bar{x}_1$		.31						
	978		.27	.24					
	693		.21	.20	.20				
	536		.21	.19	.20	.19			
	416		.20	.20	.21	.19	.14		
	326		.18	.20	.23	.25	.17	.13	
	29		.18	.26	.43	.01	.06	-.16	.36
Math 421	978 $\bar{x}_2$			.58					
	693			.54	.55				
	536			.52	.55	.50			
	416			.51	.54	.46	.41		
	326			.51	.51	.47	.39	.34	
	29			.38	.50	.31	.34	.11	.19
Math 422	693 $\bar{x}_3$				.55				
	536				.52	.49			
	416				.49	.48	.43		
	326				.46	.46	.40	.33	
	29				.54	.24	.23	.24	.18
Math 423	536 $\bar{x}_4$					.60			
	416					.56	.52		
	326					.54	.49	.47	
	29					.44	.44	.40	.43
Math 441	416 $\bar{x}_5$						.59		
	326						.57	.45	
	29						.65	.25	.48
Math 442	326 $\bar{x}_6$							.56	
	29							.62	.64
Math 443	29 $\bar{x}_7$								.21
Mean	1,244	63.4	2.2						
	978	65.3	2.4	2.3					
	693	66.3	2.5	2.6	2.1				
	536	67.4	2.6	2.7	2.3	2.2			
	416	68.8	2.8	2.8	2.5	2.4	2.3		
	326	69.1	2.9	2.9	2.7	2.5	2.5	2.4	
	29	79.1	3.4	3.3	3.2	3.2	2.9	3.2	2.2
Standard Deviation	1,244	26.4	1.2						
	978	25.8	1.1	1.2					
	693	25.8	1.0	1.0	1.2				
	536	26.2	1.0	1.0	1.1	1.1			
	416	26.1	1.0	1.0	1.0	1.0	1.2		
	326	26.0	0.9	0.9	0.9	0.9	1.0	1.1	
	29	21.9	0.8	0.8	0.8	0.8	1.1	1.0	1.2

Table 2  
Regression Coefficients and Coefficients of Multiple Correlation

Mathematics Courses	N		OSPE	Mathematics						Constant	Coefficient of Multiple Correlation
				421	422	423	441	442	443		
			$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$		
421	1,244	$y_2$	.0144**							1.26	.31**
422	978	$y_3$	.0040**	.58**						.70	.58**
423	693	$y_4$	.0025	.39**	.41**					-.10	.63**
441	536	$y_5$	.0017	.19**	.22**	.40**				.02	.65**
442	416	$y_6$	-.0004	.07	.13*	.25**	.46**			-.00	.64**
443	326	$y_7$	-.0005	.04	.03	.25**	.12	.41**		.22	.61**
601	29	$y_8$	.0116	-.28	-.06	.36	.01	.87**	-.36	-.05	.75**

\* Significantly different from zero (5% level).

\*\* Significantly different from zero (1% level).

2. Dislike mathematics—dislike some other subject ( $\chi^2 = 0.000$ ).
3. One or both parents dead—both parents living ( $\chi^2 = 0.008$ ).
4. Family fewer than 4 children—family of four or more children ( $\chi^2 = 1.543$ ).
5. Went to college the year following H. S.—additional time lapse between H. S. and college ( $\chi^2 = 4.138$ , significant at 5% level).
6. Live alone—do not live alone ( $\chi^2 = 5.250$ , significant at 5% level).

The significant values of  $\chi^2$  tend to indicate that high grades in Advanced Calculus are associated with going to college immediately after high school graduation and with rooming alone.

At the time of this writing, 25 of the 29 students had been graduated with a mean

cumulative point hour ratio of 2.89 ( $A = 4$ ). The correlation between OSPE percentiles and cumulative point hour ratio at graduation is .20, but the correlation between cumulative point hour ratio and grades in Advanced Calculus alone is .63.

### Summary

1. This study reports coefficients of correlation, means and standard deviations of mathematics course grades and Ohio State Psychological Examination percentiles.
2. Regression equations for predicting success in mathematics courses are presented.
3. Some personal data from students' official records are discussed briefly.

Received January 19, 1953.  
Early publication.

## A New Method For Determining Readability of Standardized Tests \*

Fritz W. Forbes

*University of Hawaii*

and

William C. Cottle

*University of Kansas*

Interest in the measurement of readability has been growing steadily since the first basic research was done by Vogel and Washburne (12) in 1928. Klare (8) in 1950 estimated that 34 formulas or methods for determining the reading difficulty of printed material had been devised. Five of the more recently developed formulas in widespread use have been singled out for critical analyses here: the Dale-Chall, Flesch, Lorge, Lewerenz, and Yoakam formulas.

Little has been done to ascertain the reading level necessary to understand the content of standardized testing materials. Johnson and Bond (7) have written one of the few articles on this specific topic. In their paper the Flesch formula was used for testing reading ease of nine standardized tests in common use in V. A. Advisement Centers. The general conclusion was that many tests are being administered to people who do not understand them because the readability of the tests is too difficult.

Steffle (10) made a study of the relative reading difficulty of six interest inventories using the Flesch formula. High correlation between the Flesch formula and other formulas was reported. Roeber (9) compared seven interest inventories as to word usage. The percentage of occurrence of different words appearing in the inventories was computed. He found a large number of words beyond the understanding of ninth graders. Thus, his recommendation for a glossary of terms does appear in a later form of one of the inventories.

Testing instruments are becoming so varied and numerous that persons who use them

need every help possible to determine the usefulness of the instruments for particular populations.

This study was carried out in order to determine objectively the reading difficulty of standardized tests commonly used in counseling and to develop a new and simplified method for determining the reading level of these standardized tests. It is believed that this simplified readability method will also be found useful in measuring the readability of public opinion polling questions and of headlines and slogans in advertising copy.

### Method

Five of the more popular techniques for evaluating the reading difficulty of printed matter were critically analyzed in relation to standardized tests. The Dale-Chall, Flesch, Lorge, Lewerenz, and Yoakam formulas were applied to 27 selected standardized tests commonly used for counseling at various educational levels. The mean score of reading difficulty was then obtained.

The choice of the tests to be used in this study was determined from previous studies made upon test preference and from the newer tests indicated by the records of the University of Kansas Guidance Bureau.

Berkshire and others (2) have tabulated responses that were received from 290 testing centers. They concluded that there is general agreement on approximately 15 to 20 tests as being common to guidance testing. Beyond this point test preference varies widely. Tests were chosen from this list for analysis if they were reported by at least 25 of the reporting centers as being one of the most commonly used tests. This same study shows in tabular form the results obtained

\* Abstract of Forbes' Ed.D. dissertation done at the University of Kansas under the direction of Cottle.

Table 1

Comparative Grade Placement of Selected Standardized Tests According to Various Readability Formulas and the Application of the New Forbes Formula to Items and Instructions for these Tests

Test	Dale-Chall	Flesch	Lorge	Lewerenz	Yoakam	Av. of Five Formulas	Forbes	
							Items	Instructions
MMPI	5.5	6.1	4.4	6.2	4.8	5.4	6.5	7.3
School Inventory	5.6	6.3	5.0	7.6	3.3	5.5	5.0	7.2
Calif. Test Pers.	6.2	7.2	5.3	5.7	6.6	6.2	7.1	6.3
AGCT	5.9	11.0	5.7	8.0	4.0	6.9	6.2	6.1
Guilford-Zimmerman	6.0	7.5	5.6	8.2	7.3	6.9	7.4	8.3
Otis Q-S	6.1	7.1	5.8	8.2	7.7	7.0	7.6	6.4
Adjustment Inv.	6.4	9.1	6.1	7.7	8.0	7.7	7.8	6.1
Minn. Pers. Scale	6.5	9.1	6.2	6.5	10.8	7.8	8.8	8.2
Mooney	6.1	8.3	6.0	8.7	11.0	8.1	8.9	7.2
Bernreuter	6.7	8.4	6.7	7.3	11.9	8.2	9.1	7.0
CTMM	7.2	10.8	8.8	8.0	6.7	8.3	9.1	6.3
Stanford Ach.	7.0	8.4	7.0	6.4	14.5	8.7	10.6	6.7
Kuder CM	7.3	8.5	7.7	7.8	12.5	8.7	9.7	8.3
Otis Employ.	6.3	7.9	6.3	9.3	14.3	8.8	9.1	6.4
Henmon-Nelson	6.6	9.2	6.1	11.3	12.6	9.1	9.8	5.0
Iowa Silent	8.0	11.4	7.9	9.1	10.1	9.3	9.3	6.7
Lee-Thorpe	8.0	10.0	7.9	7.8	13.6	9.5	10.3	7.6
Kuder BB	7.6	9.2	7.6	8.5	14.5	9.5	10.7	7.9
SRA Reading	8.3	13.2	8.5	9.9	12.6	10.5	9.8	6.8
Clenton	8.4	14.4	7.4	8.3	16.0*	10.9	12.5	6.9
Strong Voc. Int.	8.9	15.8	6.7	12.0	13.3	11.4	10.2	8.5
Coop. Reading	8.7	14.0	9.9	10.4	14.0	11.4	10.4	7.8
Minn. Reading	9.0	13.2	9.4	9.7	16.0*	11.5	11.9	9.2
Ohio State Psy.	10.7	16.5*	9.6	9.0	11.8	11.5	11.5	7.3
ACE	8.5	16.1	8.5	9.4	16.0*	11.7	12.7	7.6
Coop. Gen. Cult.	8.5	15.6	10.7	10.3	16.0*	12.2	Coll.	7.5
Study of Values	9.1	16.1	9.6	12.7	16.0*	12.7	12.4	9.6

\* Estimate of the grade, the formulas did not indicate grades at these levels.

from three other studies by Brophy and Long (3), Darley and Marquis (6), and Baker and Peatman (1). The findings from these three studies were comparable to those of Berkshire and others.

Standardized testing instruments that have become popular since the appearance of the above articles were checked for the frequency of their use at the University of Kansas Guidance Bureau. Nine tests were added to the original list to be analyzed. Six of these were published after the above cited studies on preference had been made. Two of the remaining three that were added were reading tests, because of the nature of the present study. The one remaining test, the *Minne-*

*sota Personality Scale* (Men) 1941, was added at the discretion of the writers. The tests were chosen from five of the general areas of testing listed by the *Third Mental Measurements Yearbook* (4): Character and Personality, Intelligence (group), Interests, Achievement Batteries, and Reading. They are listed in Table 1.

The five formulas selected for study are the more recently developed techniques for measuring readability. They present several factors which have been used for determining reading difficulty of printed matter, namely, word difficulty, prepositional phrases, sentence length, number of syllables per one hundred words, number of different words

and percentage of words beginning with certain letters. Each of the formulas has been carefully developed and exhibits a fair degree of reliability and validity. In the Yoakam formula, "hard words" vary in difficulty according to their frequency and range of occurrence above the most common four thousand words. The Lewerenz formula is based solely on word difficulty, basing the vocabulary difficulty on words with certain initial letters. The Flesch formula considers the length of the word the index of difficulty of that word, the more syllables a word has the more difficult it is. The Dale-Chall formula uses a list of three thousand words, any word not appearing on this list is considered difficult. The Lorge formula considers as a "hard word" any word other than the 769 words that are common to the first one thousand most frequent English words on the Thorndike list and the first thousand most frequent words known by children entering the first grade.

On the basis of the facts mentioned above, it would seem that the Lorge formula interpretation of "hard words" is too simple and limited for the purpose outlined here; the Dale-Chall method approaches a more realistic and practical definition of difficult words; and the Yoakam formula is perhaps the most realistic of all the formulas for use with testing instruments. The idea that difficult and easy words begin with certain letters, presented in the Lewerenz formula, does not seem to apply to standardized tests; and the number of syllables that a word has, as proposed by Flesch, does not necessarily give its index of difficulty.

The grade level scores obtained for each test from the five formulas were averaged in order to obtain a mean grade level reading difficulty score for each test. These mean scores were taken as criterion grade level scores of reading difficulty for these selected tests. They are shown in Table 1.

A definite difference was noted in the results of the measurement of the various tests by these five formulas as shown in Table 1. There was as much as 8.13 grades difference in the reading difficulty of a single test as determined by two different formulas.

At the same time the five formulas correlated significantly with each other. The rank order correlations ranged from .91 between the Dale-Chall and Flesch formulas to .59 between the Lewerenz and Yoakam formulas. These intercorrelations are shown in Table 2.

The rank order correlations between each of the formulas and the mean grade level score ranged from .95 for the Dale-Chall formula to .77 for the Lewerenz formula as shown in Table 2.

Correlations were also computed by means of the ratio of the estimated true variance to the observed variance between each formula and the means of the five formulas (5). These correlations as shown in Table 3 ranged from .90 between the Dale-Chall formula and the mean of the five formulas to .84 between the Flesch formula and the mean of the five formulas. However, the scores obtained from this study for the reading level of each test correlated slightly over .95 with the mean of the five formulas and ranged between .95 and .72 for the five formulas as shown in Table 3.

Table 2  
Intercorrelations (Rho) for the Five Formulas Applied to Twenty-seven Tests and Correlation (Rho) Between Each Formula and Mean of the Five Formulas

Formula	Dale-Chall	Flesch	Lorge	Lewerenz	Yoakam	Mean of Five
Dale-Chall	—	.91	.90	.65	.75	.95
Flesch			.81	.66	.66	.90
Lorge				.60	.69	.89
Lewerenz					.59	.77
Yoakam						.89
Mean of Five						—

Table 3

Correlations Between Forbes' Formula, the Five Formulas Applied to Twenty-seven Tests and the Mean of the Five Formulas Using Ratio of Estimated True Variance to Observed Variance (5)

Formula	Dale-Chall	Flesch	Lorge	Lewerenz	Yoakam	Mean of Five
Forbes	.95	.83	.84	.72	.90	.96
Mean of Five	.90	.84	.87	.86	.87	

The knowledge of grammar needed to apply most of these five formulas studied is considerable. Also, the amount of time required by these methods makes them quite laborious. More than ten hours were required to apply some formulas to a single test. The average amount of time for the working of a single formula on a single test was more than two and one half hours. The simplified Forbes method, in contrast, requires only approximately one half hour per test.

Word difficulty was used as a common factor in all five formulas studied. It was also evident from a review of the literature that word difficulty was basic to the readability of all printed matter.

#### Development of the Forbes Method

The following steps were taken in developing the Forbes method which is specifically suited for measuring the readability of printed matter in standardized tests:

1. The five formulas studied were applied to each of the twenty-seven standardized tests.

2. The mean grade level score of the five formulas for each test was taken as the criterion of readability for the tests.

3. The vocabulary difficulty was determined for each test by finding the number of words above the most frequently used 4,000 words in three samples of 100 words each selected at the beginning, middle, and end of each test.

4. The *Thorndike Junior Century Dictionary* was used for finding the weights to be assigned to each word above the most commonly used four thousand. The number following the definitions in this dictionary is the weight for that word. The weights range from one to twenty, but since the first four

thousand were dropped, only numbers of four and above were used.

5. The total of these weights for each test was divided by the number of words in the samples, giving the index of vocabulary difficulty.

6. The standardized tests studied were placed in rank order as determined by the mean grade level scores of the five formulas. These tests were set off into grade groups. All tests falling within one half grade level above or below the grade were considered with that grade group. For example, grade level scores of 7.5 to 8.5 would be considered characteristic of the eighth grade reading difficulty. The largest and smallest indices of vocabulary difficulty falling within any one grade group were considered the limits for that grade. Table 4 gives the indices of vocabulary difficulty, setting the limits for the various grade levels.

7. The grade level scores derived from this method give the average reading grade level required for the person taking the test in

Table 4

Grade Level of Reading Difficulty as Determined by the Index of Vocabulary Difficulty

Index of Vocabulary Difficulty	Grade Level
1.4510 and above	College
1.2510-1.4509	12th grade
1.0510-1.2509	11th grade
.8510-1.0509	10th grade
.6510-.8509	9th grade
.4510-.6509	8th grade
.2510-.4509	7th grade
.0510-.2509	6th grade
.0509 and below	5th grade

order that the test be understood and done properly.

The exact method of applying the new formula is listed as follows:

1. Three samples of 100 words each were taken from the tests to be analyzed. The samples were selected in each test at the beginning, middle, and end. The only requirements for the samples were that they consist of an even hundred words, that each sample begin with the first word of an item, and the vocabulary tests be omitted from the samples. It seemed only fair to omit the vocabulary sections in order to get the average reading difficulty of the standardized tests.

It seemed easiest to begin with the first word of the first item of a test and count the first hundred word sample exactly. The middle sample was selected as near the midpoint of the test as possible. Starting with the middle item count backward to the initial word of an item close to fifty words back. The remainder of the middle one hundred word sample was secured by counting the difference from one hundred in words beyond this middle item. The third sample was taken by counting backwards from the last word of the test items until one hundred words were counted. Should the one hundred words end within the item, proceed counting backwards until the first word of an item is reached, then in order to get exactly the one hundred words omit the number over one hundred at the end of the sample.

2. Each word that appeared difficult to the grader was written on a sheet of paper. These words were then found in the 1942 *Thorndike Junior Century Dictionary* (11). The number following the definition in this dictionary is the weight for that word. These numbers range from one to twenty, representing the first twenty successive thousands of words most commonly used in the English language. Only words above the most frequently used four thousand words were given a weight. Any word having a weight of four or above was considered a difficult word and its weight was listed. Words used more than once in the samples were given their weights *each* time they were used.

3. The weights for the three samples were

totaled and divided by the number of words in the samples, 300 in this case. This gave the *index of vocabulary difficulty* for the standardized tests.

4. Using the indices of vocabulary difficulty obtained from the above three steps, refer to Table 4 in order to determine grade level of difficulty of the printed matter in the test being analyzed. Grade level scores may be interpolated to the nearest tenth of a grade.

5. The reading difficulty was also figured for the instructions for each of the tests analyzed. The samples in some cases included all directions to the tests when they consisted of 300 words or less; other samplings followed the procedure outlined above for the test, that is, taking 100 word samples at three points throughout the instructions.

There is little room for decisions to be made by the scorer who uses the Forbes method since the words are weighted in accordance with an accepted word list. If a variant of a word or a hyphenated word does not appear in this list, no weight is given. Only words that appear in the *Thorndike Junior Century Dictionary* are given weights.

#### Summary

1. Review of the literature showed that no specific method has been developed for finding the reading difficulty of standardized tests (or public opinion polling questions or headlines and slogans in advertisements) up to the present time.

2. The five techniques for measuring the readability of printed matter that were applied to the 27 standardized tests in this study showed wide variation as to the grade placement of the reading difficulty of these tests.

3. The usual methods in use for determining the readability of reading material consume a great amount of time for their application.

4. These methods also required much interpretation and judgment on the part of the user, thus greatly lessening their objectivity.

5. The peculiar make-up of the reading matter in standardized tests required that only the vocabulary difficulty factor be used for

determining their readability. The use of such factors as sentence length and prepositional phrases was not practical since many of the tests have sections composed only of word lists.

6. The instructions to the standardized tests were easily within the range of reading difficulty of those for whom the tests were designed.

7. The use of short word lists for determining difficult words tended to give too coarse a classification of grade levels of reading. A longer list made the method for determining the readability of standardized tests more sensitive, spreading the grade level scores over a longer range.

8. The method developed in this study was based entirely upon reading matter found in commonly used standardized tests. It is a technique applicable only to such reading matter or to similar material.

9. The method evolved in this study is easily applied, consumes little time, and shows high objectivity by the elimination of most of the interpretations and judgments formerly left to the scorer.

Received August 4, 1952.

#### References

1. Baker, G., and Peatman, J. G. Tests used in Veterans Administration advisement units. *Amer. Psychol.*, 1947, 2, 99-102.

2. Berkshire, J. R., Bugental, J. F. T., Cassens, F. P., and Edgerton, H. A. Test preference in guidance centers. *Occupations*, 1948, 26, 337-343.
3. Brophy, D. F., and Long, L. Veterans Administration vocational training program: Processing procedures used by the College of the City of New York. *Psychol. Bull.*, 1944, 41, 795-802.
4. Buros, O. *The Third Mental Measurements Yearbook*. New Brunswick: Rutgers University Press, 1949.
5. Cottle, W. C. Card versus booklet forms of the MMPI. *J. appl. Psychol.*, 1950, 34, 255-259.
6. Darley, J. G., and Marquis, D. G. Veterans guidance centers: A survey of their problems and activities. *J. clin. Psychol.*, 1946, 2, 109-116.
7. Johnson, R. H., and Bond, G. L. Reading ease of commonly used tests. *J. appl. Psychol.*, 1950, 34, 319-324.
8. Klare, G. R. *Evaluation of quantitative indices of comprehensibility in written communication*. Unpublished Ph.D. Thesis, University of Minnesota, 1950.
9. Roeber, E. C. A comparison of seven interest inventories with respect to word usage. *J. educ. Res.*, 1948, 42, 8-17.
10. Steffire, B. The reading difficulty of interest inventories. *Occupations*, 1947, 26, 95-96.
11. Thorndike, E. L. *Thorndike Junior Century Dictionary, Revised Edition*. New York: Scott, Foresman and Company, 1942.
12. Vogel, M., and Washburne, C. An objective method of determining the grade placement of children's reading material. *Elem. Sch. J.*, 1928, 28, 373-381.

## A Modified Administration Procedure for the O'Connor Finger Dexterity Test

Edwin A. Fleishman

*USAF Air Training Command, Human Resources Research Center \**

The O'Connor Finger Dexterity Test (8) has been widely used in counseling and selection.<sup>1</sup> It appears to measure dexterity of a finer type than is measured by the Minnesota Rate of Manipulation Tests (11, 9), or by most of the subtests of the Purdue Pegboard (12). The test seems most useful for manual jobs requiring rapid wrist and finger movements, in fine assembly work requiring both speed and precision, and in jobs involving rapid manipulation of small objects. The validity of the test has on occasion been demonstrated for electrical fixture assemblers (14, 16), radio assemblers (16), power-sewing machine operators (10), watch assemblers (1, 3), punch press operators (7), can packers (13), and dental students (5, 6).

Despite its widespread use, the test has two primary difficulties as a selection device. First, relative to other dexterity tests (e.g., Minnesota Rate of Manipulation, Purdue Pegboard), the test takes considerably longer to administer. The time required generally varies from 8 to 15 minutes. Moreover, during this more lengthy time period, the test yields only one score, whereas in a considerably shorter administration time the Purdue Pegboard yields five scores (right, left, both hands, total of these, assembly), and the Minnesota Rate of Manipulation Test yields at least two scores (placing and turning).

A second limitation of the O'Connor test for selection purposes is that it is a work limit test. The examinee's score is the total number of seconds it takes him to fill the board.

\* Perceptual and Motor Skills Research Laboratory, Lackland Air Force Base, San Antonio, Texas. The data reported in this study were collected as part of the United States Air Force Human Resources Research and Development Program. The opinions or conclusions contained in this report are those of the author. They are not to be construed as reflecting the views or indorsement of the Department of the Air Force.

<sup>1</sup> This paper is *not* concerned with the O'Connor Tweezer Dexterity Test, which has also received considerable study.

This procedure makes it difficult for a single examiner to administer the test to more than one subject at a time.

The present study investigated the feasibility of certain modified administration procedures which would decrease the total time required to give the test and which would render the test more suitable for group administration.<sup>2</sup>

### Procedure

The O'Connor Finger Dexterity Test was administered under *time limit* conditions to unselected samples of basic airmen at Lackland Air Force Base. The mean age of the subjects was 18.9 with a standard deviation of 1.3. The test was administered to independent samples of 100 subjects each. One group received the test for a four-minute time limit condition, another group for a five minute period, and a third group for a six-minute period. Within each sample the tests were re-administered for test-retest reliabilities. The interval between test and retest was held constant at one and one-half hours for each group, since it was assumed that the length of the interval might affect the magnitude of the reliability coefficients. Under these time limit conditions, a subject's score was the total number of pins placed during the allotted time.<sup>3</sup>

In another sample, 100 subjects were tested and retested one and one-half hours later under the standard *work limit* conditions in which the total number of seconds required

<sup>2</sup> There appears to be very little published evidence indicating whether or not time limit and work limit methods of administering speed tests are equivalent and interchangeable. In one of the few previous studies on this problem, Paterson and Tinker (11a), working with speed of reading tests, found the work limit method to agree with the time limit method as closely as each method agreed with itself.

<sup>3</sup> Since 3 pins are placed in each hole the subject's score is three times the number of holes filled up to the last hole plus the number of pins in the last hole. There are 100 holes in the total board.

Table 1  
Means, Standard Deviations, and Reliabilities for Five Administration Conditions<sup>1</sup>

Testing Method	Score	Test		Retest		Reliability
		M	S.D.	M	S.D.	
Work limit, full board	Seconds	556	62.8	502	53.3	.86
Work limit, half board	Seconds	277	34.8	256	27.0	.82
Time limit—6 minutes (360 seconds)	Number of pins	205	19.9	221	23.9	.80
Time limit—5 minutes (300 seconds)	Number of pins	166	19.9	178	19.4	.76
Time limit—4 minutes (240 seconds)	Number of pins	137	14.9	147	14.1	.71

<sup>1</sup> Data for the Work Limit tests are based on the same sample of 100 Airmen. Data for each Time Limit test are based on separate samples of 100 each.

to fill the entire board was recorded.<sup>4</sup> Also recorded during these administrations was the time required to fill half the board. In an additional sample, 100 subjects were given the test under work limit conditions and were re-tested one and one-half hours later under the five minute time limit condition.

In all, 500 subjects were involved in the study. Independent groups were used in each phase in order to duplicate the standard testing conditions. In this way scores and reliability coefficients derived from each administration procedure could be more readily compared, uncomplicated by differential practice effects from other forms of the test.

### Results

Table 1 presents the means, standard deviations and test-retest reliabilities for the various administration procedures.

It should be noted that these reliability coefficients are to be regarded as conservative relative to split-half or immediate retest reliability estimates often reported. For example, Darley (4) has reported a corrected split-half reliability of .90 and Blum (1) reported a test-retest reliability of .89 for the standard test with a half-hour interval between administrations. This latter reliability compares

favorably with our test-retest reliability of .86 following a longer (one and one-half hours) interval. He also reports a reliability of .82 for the half length test which is identical with our results. All these coefficients are higher than the original test-retest reliability of .60 reported by Hines and O'Connor (8) in their original standardization of the test.

The correlations obtained between the time limit procedure (five minutes), and the full board and half board work limit procedures were .96 and .89, respectively, after correction for attenuation. This gives some indication that the abilities measured by the time limit and work limit forms of the test are the same.

It can be seen in Table 1 that there is some loss in reliability when the test is administered as a time limit test. In order to achieve comparable reliability under time limit conditions, to that obtained in the full work limit test (.86), nine minutes testing time would probably be required.<sup>5</sup> However, the reliability achieved in the six-minute trial is probably sufficient for group prediction purposes. If one is using the test as part of a larger selection battery, one of the shorter tests probably has sufficient reliability for inclusion, since a reduction in reliability of this magnitude would have little effect on the composite validity of the battery (see 2, 15). The four-minute test might well be used where a choice

<sup>4</sup> The original method of scoring the test involved a small correction in the second half of the test for practice on the first half. However, Tiffin and Greenley (16) found a correlation of .99 between the total time score and scores obtained by the original formula. More recent studies, including those of the USES, have used the simpler total time score.

<sup>5</sup> Estimated from the six-minute test by the Spearman Brown prophecy formula.

Table 2

Normative Data for Five Administration Conditions of the O'Connor Finger Dexterity Test<sup>1</sup>

Percentile	Raw Scores				
	Conditions				
	Work Limit (Seconds)		Time Limit (No. of pins placed)		
	Full Board	Half Board	6 Min.	5 Min.	4 Min.
100	372	213	270	264	180
99	430	217	245	219	173
98	442	221	242	214	169
96	468	226	236	208	165
90	481	233	228	196	157
75	507	254	216	185	146
63	526	262	210	178	139
50	548	271	204	170	135
37	572	278	198	166	131
25	600	299	189	157	127
14	632	317	182	147	124
10	647	323	178	142	122
6	674	346	174	136	116
2	717	379	168	119	113
1	746	384	140	115	101

<sup>1</sup> Norms for the Work Limit tests are based on a sample of 200 Airmen. Norms for each Time Limit test are based on separate samples of 100 each.

must be made (as is often necessary) between including a longer form of the test or adding some additional type of test which broadens the scope of abilities sampled by the battery in the time allowed. For individual prediction and guidance purposes, the standard work limit procedure is probably desirable.

Table 2 summarizes some preliminary normative data for the various administration procedures.

Although these results are based on limited samples and are to be regarded as specific to this kind of population, they may serve as a suggestive guide in future use of the test under these conditions. It is also to be noted that the work limit scores presented are generally higher (poorer performance) than those usually reported for other populations.

### Summary

The O'Connor Finger Dexterity Test was administered under various work limit and short time limit conditions. The results in-

dicade that although there is some loss in reliability under the time limit conditions, the reliabilities are probably adequate for group prediction, especially if the test is to be included in a larger battery. Preliminary norms for the modified administration conditions were presented.

Received August 12, 1952.

### References

1. Blum, M. L. A contribution to manual aptitude measurement in industry. *J. appl. Psychol.*, 1940, 24, 381-416.
2. Brokaw, L. D. Comparative validities of "short" versus "long" tests. *J. appl. Psychol.*, 1951, 35, 325-330.
3. Candee, B., and Blum, M. L. Report of a study done in a watch factory. *J. appl. Psychol.*, 1937, 21, 572-582.
4. Darley, J. G. Reliability of tests in the standard battery (in "Research studies in individual diagnosis"), Univ. of Minnesota, *Bull. Empl. Stab. Res. Inst.*, No. 4, 1934.
5. Douglass, H. R., and McCullough, C. M. Prediction of success in the School of Dentistry. *Univ. Minn. Stud. Predict. School Arch.*, 1942, 2, 61-74.

6. Harris, A. J. Relative significance of measures of mechanical aptitude, intelligence, and previous scholarship for predicting achievement in dental school. *J. appl. Psychol.*, 1937, 21, 513-521.
7. Hayes, E. G. Selecting women for shop work. *Personnel J.*, 1932, 11, 69-85.
8. Hines, M., and O'Connor, J. A measure of finger dexterity. *J. Personnel Res.*, 1926, 4, 379-382.
9. Jurgensen, C. E. Extension of the Minnesota Rate of Manipulation Test. *J. appl. Psychol.*, 1943, 27, 164-169.
10. Otis, J. L. Prediction of success in power sewing machine operating. *J. appl. Psychol.*, 1938, 22, 350-366.
11. Paterson, D. G., and Darley, J. G. *Men, women, and jobs*. Minneapolis: Univ. of Minnesota Press, 1936.
- 11a. Paterson, D. G., and Tinker, M. A. Time-limit and work-limit methods. *Amer. J. Psychol.*, 1930, 42, 101-104.
12. Purdue Pegboard (Examiner Manual). Science Research Associates, 228 South Wabash, Chicago, Illinois.
13. Stead, W. H., and Shartle, C. L. *Occupational counseling techniques*. New York: American Book Co., 1940.
14. Steel, M., Balinsky, B., and Lang, H. A study on the use of a work sample. *J. appl. Psychol.*, 1945, 29, 14-21.
15. Thorndike, R. L. *Personnel selection*. New York: John Wiley, 1949.
16. Tiffin, J., and Greenly, R. J. Employee selection tests for electrical fixture assemblers and radio assemblers. *J. appl. Psychol.*, 1939, 23, 240-263.

# A Comparison of the Revised Allport-Vernon Scale of Values (1951) and the Kuder Preference Record (Personal)

Ira Iscoe and Omer Lucier

University of Texas

Both the Allport Vernon Scale and the Kuder Preference Record (Personal) yield separate "trait" scores which are named and defined. The purpose of the research herein reported was to examine the communality of the various trait scores on the two scales. According to the definitions given in the respective manuals (1, 3), it would be expected that a high positive correlation would exist between: (1) the Theoretical "trait" scores of both instruments; (2) the Economic "trait" scores of the Allport and the Practical of the Kuder; and (3) the Political of the Allport and the Sociable and the Dominant of the Kuder.

It would be also expected that a high negative correlation would exist between the Aesthetic of the Allport and the Theoretical of the Kuder.

## Subjects

Ninety adult males, the majority of them University of Texas students, acted as sub-

jects. The mean age of the group was 26 years with an S.D. of 6.5 years. The average number of years of education was 14.6 with an S.D. of 2.4. The mean scores and S.D. made by the experimental groups were not significantly different from the scores made by comparable groups used by Allport and Kuder in standardizing their tests.

## Procedure

The tests were administered in accordance with the instructions in the respective manuals. The Allport-Vernon was taken first, followed by the Kuder. If the subjects did not have time to complete the Kuder during the scheduled sessions it was taken home and returned later. Scattergrams were made for the numerous combinations of item scores for one inventory with item scores of the other inventory. Since a rectilinear relationship was obtained for all scattergrams the use of the Pearson product-moment formula for correlation was justified. The evaluation of the data

Table 1  
Correlations Between Each Score of the Kuder Preference Record (Personal)  
and Each Score of the Allport-Vernon Scale of Values, 1951

Allport	$r^*$	Kuder				
		Sociable $r^* = .86$	Practical $r^* = .86$	Theoretical $r^* = .85$	Agreeable $r^* = .84$	Dominant $r^* = .85$
Theoretical	(.87)	-.30	.04	.20	-.08	.01
Economic	(.92)	.00	.09	-.52	.16	.13
Aesthetic	(.90)	.01	-.23	.47	.18	-.08
Social	(.77)	.10	.05	.17	.10	-.08
Political	(.90)	.02	.33	-.36	-.32	.30
Religion	(.90)	.13	.27	-.08	.13	-.16

\* The score reliabilities are from the respective manuals, and are placed in parentheses immediately following the Allport designation and immediately underneath the Kuder designation. The Kuder manual contains reliability measures computed by The Kuder-Richardson Formula. The population selected for use in Table 1 was that for "100 men." The reliabilities from the manual for the Allport scores were obtained from "Test-Retest" data for 34 cases with one month intervening between the test and the retest. \*The manual also includes a table of split-half reliabilities with an N of 100. Statistical material on the revised form of this inventory is as yet scarce due to the recency of its publication (1951). According to the authors, "the present revision offers certain improvements without in any way changing the basic purpose of the test (referring to the 1931 version of the 'Study of Values' as compared to the 1951 version) or limiting its scope of usefulness" (1, p. 6).

by means of factor analysis was not resorted to in view of Guilford's (2) recent article on "Ipsative" factors and his remarks that scales such as the Kuder were not amenable to factor analysis. A total of 30 correlations (six traits of the Allport and five for the Kuder) were computed.

### Results

It can be seen from Table 1 that none of the hypotheses put forth at the beginning of this article were justified. The correlation of .20 between the two theoretical scales is surprisingly low. Indeed, the highest positive correlation obtained (.47) was between the aesthetic of the Allport and the theoretical of the Kuder—where the expectancy was for a high negative correlation. The low positive correlation between the "Social" of the Allport and the "Sociable" of the Kuder can be explained in that other than having similar names, they are defined rather differently.

### Conclusions

The results obtained point up once again the dangers of using similarly defined traits measured by different instruments. As an example, one of our subjects obtained the following raw scores on the "Theoretical" of both instruments:

Instruments	Raw Score	Percentile Rank
Allport	48	73
Kuder (Personal)	25	13

It can be seen that on one instrument he would be considered of reasonably high theoretical orientation while on the other he would be very low. Since both the Allport and the Kuder are used in educational and vocational counseling a totally different picture of this subject's interest would have been furnished. The importance of knowing the relationships between the various measuring instruments is perhaps one way of avoiding gross errors in the counseling situation. One avenue of further research might be the use of two instruments on a population where certain traits mentioned were believed to be present to a high degree.

*Received July 17, 1952.*

### References

1. Allport, G. W., Vernon, P. E., and Lindzey, G. *Manual to the Study of Values*. New York: Houghton Mifflin Co., 1951.
2. Guilford, J. P. When not to factor analyze. *Psychol. Bull.*, 1952, 49, 26-27.
3. Kuder, G. F. *Manual for the Kuder Preference Record (Personal)*. Chicago: Science Research Associates, 1949, Revised Sept. 1949.

## Administering Form BB of the Kuder Preference Record, Half Length

A. A. Canfield<sup>1</sup>

Wayne University

To administer Form BB of the Kuder Preference Record<sup>2</sup> to an entire group of people in industry usually requires more than an hour although many finish in less time. Feelings of fatigue, boredom, and annoyance are commonly expressed by the examinees during the course of the test. Sighs of relief are regular expressions for employed adults who complete the record. The laborious and frequently painful task of punching the holes with the small pin provided, the turning of progressively smaller and smaller pages, and the apparent duplication of items from page to page combine to give an emotional reaction that is, in the main, unpleasant. Examinees often ask if some scoring technique is used to "check-up" on the consistency of their answers by comparing their answers on what they believe to be the same item occurring on different pages.

The objections raised by the examinee are not always easy to turn away with good conscience, for the specific percentile scores obtained are normally sorted into quite broad and often arbitrary categories for interpretation. The manual<sup>3</sup> accompanying the test in the section devoted to the interpretation of the section recommends that three general score categories be used in interpreting the results (high for percentiles above 75, low for percentiles below 25, and average for the middle 50%). Some test users have expanded this to include five groupings. It can be very embarrassing to try to explain to an examinee

the need for the length of the test, if one admits to this type of interpretation, an interpretation which many users of the test make. In addition to these broad classifications, the manual contains sample profiles for 51 occupations. The manual cites the desirability of collecting more data for the purpose of preparing occupational profiles of greater reliability, but does not recommend their use in counseling or guidance work.

In many cases a measure of interests, such as this test provides, would be a useful adjunct to the information supplied by other tests but the testing time required makes it impractical. Miles<sup>4</sup> has recognized the problem and suggested using the scores obtained on pages 7, 8, and 9 of the record and then multiplying them by constants to predict the total score on each of the nine interest areas. Using this method on a sample of 205 adult males, correlations were obtained between the predicted and the actual scores ranging from .76 on Part III (Scientific) to .91 on Parts II and VI (Computational and Literary).

The present study was undertaken because it was considered desirable to elaborate this ratio approach by making a correlational analysis and developing regression equations for a more accurate prediction of the total score. An examination of the answer sheet showed that the division of pages in the booklet that comes the closest to giving an even division of item responses for the nine interest areas was that of the odd-numbered pages versus the even-numbered pages. Since this division also supplied an odd-even reliability grouping, it was decided to undertake the study using this page division.

### Method

A total of 301 completed records, representing a substantial proportion of the per-

<sup>1</sup> The author wishes to express his thanks to the firm of George Fry and Associates in Chicago, Illinois for making these data available for the research and for providing facilities and supplies for the processing of the data, and to Mr. Wesley Potter of Northwestern University for his conscientious application to the laborious task of scoring most of the papers and preparing frequency distributions.

<sup>2</sup> Kuder, G. F. *Kuder Preference Record, Form BB*. Chicago: Science Research Associates, 1942.  
<sup>3</sup> Kuder, G. F. *Revised Manual for the Kuder Preference Record*. Chicago: Science Research Associates, 1946.

<sup>4</sup> Miles, R. W. A proposed short form of the Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 282-285.

Table 1  
Correlations and Regression Equations Obtained in the First Sample  
N = 301

Interest Area	Odd Pages		Odd-Even		Even Pages	
	$r_{et}$	Regression Eq.	$r_{oe}$	$r_{oe}^c$	$r_{et}$	Regression Eq.
1. Mechanical	.96	$1.38x + 20.18$	.92	.96	.89	$1.91x + 9.10$
2. Computational	.92	$1.85x + 3.30$	.77	.87	.93	$1.68x + 4.04$
3. Scientific	.92	$1.82x + 3.95$	.78	.88	.93	$1.68x + 11.35$
4. Persuasive	.94	$1.81x + 7.60$	.84	.91	.94	$1.83x + 7.97$
5. Artistic	.92	$1.65x + 1.02$	.78	.88	.95	$1.99x + 5.93$
6. Literary	.92	$2.01x + 8.28$	.81	.90	.92	$1.60x + 3.31$
7. Musical	.88	$1.83x + 2.20$	.78	.88	.91	$1.74x + 1.63$
8. Social Service	.93	$1.80x + 5.36$	.76	.86	.92	$1.67x + 13.97$
9. Clerical	.97	$2.10x + 6.29$	.69	.82	.91	$1.38x + 9.04$

sons employed in supervisory, staff and advisory, and skilled line positions by a large midwestern canning company, were pulled from the files. Each paper was then scored by interest area, the score on each area being divided into that achieved on the odd-numbered pages and that obtained on the even-numbered pages of the booklet. Inasmuch as these scores, as well as the totals in some cases, were noticeably skewed, all of the distributions were normalized by the percentage method. Correlations were then computed between each of these scores and the other two for each interest area. The computation of the mean and standard deviation of each distribution supplied the additional data necessary to develop the regression equations desired for predicting the total scores from the scores on either of these two halves.

To check upon the accuracy of these equations the completed records of a second sample of 100 employed males, drawn alphabetically from the files, were scored in the same manner. The score in each interest area was broken down into that obtained on the odd-numbered pages and that achieved on the even-numbered pages. None of the papers used in the first sample were included in this second group. Correlations were then computed between the predicted scores and the obtained scores for each of the nine interest areas, and the standard errors of estimate obtained. As a check upon the representa-

tiveness of the two samples used in this research, the means and standard deviations of the total scores in each of the interest areas were also computed.

### Results

The original correlations between the scores on the even-numbered pages and the total scores, and the resulting regression equations are presented in the first two columns of Table 1. The correlations between the scores on the odd-numbered and even-numbered pages for each of the nine interest areas are shown in the third column. Inasmuch as they represent odd-even reliability figures, the corrected values, using the Spearman-Brown prophecy formula, are shown in the adjacent column. These values are almost identical with those given in the manual for reliabilities computed using the Kuder-Richardson method. The two right hand columns of Table 1 show the correlations between the scores on the even pages and the total scores, and the resulting regression equations. These regression equations were then used for predicting the total scores as previously described.

The correlations obtained between the predicted total scores, based on the odd-page scores, and the obtained total scores for the second sample, along with the mean errors of prediction and the standard errors of estimate are shown in the first three columns of Table 2. The same information for the predictions

Table 2

Correlations Between Predicted and Obtained Scores, Mean Errors, and Standard Errors  
of Estimate Obtained in the Second Sample  
N = 100

Interest Area	Odd Pages			Even Pages		
	$r_{yy'}$	$M_{error}$	$S.E._{est}$	$r_{yy'}$	$M_{error}$	$S.E._{est}$
1. Mechanical	.97	1.32	4.94	.93	.21	7.54
2. Computational	.94	-.95	4.06	.93	.10	4.17
3. Scientific	.92	.49	5.51	.94	-.48	5.08
4. Persuasive	.97	-.54	5.39	.96	-.40	5.87
5. Artistic	.95	-.85	4.27	.94	.44	4.63
6. Literary	.92	.95	4.96	.96	-.65	3.73
7. Musical	.90	.16	3.16	.93	-.22	2.70
8. Social Service	.91	.16	6.18	.92	-.32	5.79
9. Clerical	.92	-.33	5.39	.96	-.03	3.99

based on even-page scores are given in the last three columns of Table 2.

It will be noted that the correlations between the predicted total scores and the obtained total scores range from .90 to .97. The standard errors of measurement are small, considering the percentile equivalents, and the mean errors similarly small.<sup>5</sup>

Table 3 shows the means and standard deviations of the two groups used in this study and the means and standard deviations

<sup>5</sup> Conversion tables have been prepared which make it possible to translate the part-score directly into the percentile score in each of the nine interest areas. A copy of these conversion tables can be secured from the Department of Personnel Methods, School of Business, Wayne University, Detroit 1, Michigan at no cost.

of the norm group reported in the manual. The means and standard deviations of the three groups are generally similar, with the exception of the generally higher interests of the experimental groups in the persuasive area.

### Summary

This study was designed to determine the plausibility of administering Form BB of the Kuder Preference Record half length. An analysis of the test answer sheet indicated that the odd pages and the even pages of the test contained a fairly even distribution of items in each of the nine interest areas measured by the test.

Table 3

Means and Standard Deviations of the Two Sample Groups and the Test Norm Group

Interest Area	Prelim. Study (N = 301)		Verification Study (N = 100)		Norm Group (N = 2667)	
	M	S.D.	M	S.D.	M	S.D.
1. Mechanical	81.3	18.8	74.3	20.3	78.6	22.8
2. Computational	36.8	11.4	35.0	11.7	35.3	10.6
3. Scientific	63.2	14.8	61.8	14.5	64.0	15.5
4. Persuasive	86.6	20.9	92.4	21.0	74.4	20.6
5. Artistic	44.7	12.9	40.6	14.0	46.1	13.6
6. Literary	49.8	14.3	50.0	12.7	47.8	15.1
7. Musical	16.1	7.8	17.1	7.2	16.6	9.6
8. Social Service	73.6	16.8	74.6	15.1	73.7	17.5
9. Clerical	48.0	13.3	47.8	14.1	52.1	13.5

The completed answer sheets of 301 employed males, representing a variety of jobs, were analyzed and mathematical equations for predicting total scores from the scores on the two different sets of pages were developed. A second sample of 100 employed males was used to test the accuracy of predictions using these equations.

The results indicate that the test could be administered half length with little loss of accuracy under normal conditions of test interpretation. This reduction in administra-

tion time means an appreciable reduction in testing costs, greatly lessened feelings of fatigue, boredom, and irritation for the examinee, greater possibilities for using the test in industrial situations where its administration time has previously been considered prohibitive, and an opportunity to use the time saved for the administration of other tests that might contribute to the prediction of job success.

*Received June 30, 1952.*

## Attitudes Toward Public Low-Rent Housing, Before and After Construction \*

Kenneth E. Clark

*University of Minnesota*

and

Charles E. Swanson

*Institute of Communications Research, University of Illinois*

Two years prior to the collection of the data reported herein, there was announced in a midwestern city a plan for the erection of a public housing project, using federal funds, to provide living facilities for persons of low income. Since this project was to be built in an area fairly well surrounded by existing housing, it was considered that the survey of attitudes of persons in the neighborhood both before and after construction of the development would provide significant information on the dynamics of attitude changes. The results of the original survey before construction have already been reported in this journal.<sup>1</sup>

The two surveys were made under fairly comparable conditions. The first was made in June 1950, shortly after announcement of approval of the project. The second survey was made just two years later, approximately one year after construction had started, and about two weeks before the first families began to move into the project. The first sample was drawn as a fixed-address City Directory sample; of 196 units listed, interviews with a responsible adult were obtained in 188, or 96 per cent. In the second sample this same list of addresses was used, supplemented by an additional sample of 192 residences. A total list of 388 addresses was obtained, of which 366 were usable (17 had been torn down, 4 addresses were erroneous, 1 house was vacant). Of these 366, 351

households, or 96 per cent, were contacted and interviewed. Six householders refused to be interviewed (less than 2 per cent); 2 would not answer the door; 2 were not at home after 2 call-backs; 2 were out of town; 2 returns were not usable; 1 was unable to help because of death in the family.

The same questionnaire used prior to construction was used after construction, with only minor changes ("how many stories *will* . . ." was changed to "how many stories *does* . . ."). A new question was added to permit sorting the householders into those present in the community at the time of the first survey and those not present.

### Results

Opposition to the housing project decreased somewhat over the two year period. Responses to the questions "Do you favor or oppose the construction of this new development?" and "How strongly do you feel about this?" for the two years, 1950 and 1952, are presented in Table 1. This increase in favor occurs without a reduction in the number of no opinion responses. This is a rather surprising result, since one might expect that, with the project in actual physical existence, more persons would have formulated an opinion of some sort. The physical appearance of these new units in general attracted favorable comment, which may account in part for the shift in attitude, but certainly not for the continued high percentage of persons refusing to state a position.

The same question asked about income in 1950 was asked again in 1952, except that an additional \$500 was added to each response category as a rough estimate of the average increment in income which might have been

\* The writers are indebted to Mr. Norris Ellertson for his work in the supervision of the interviewing staff used in this study, and for his work in the analysis of results. This study was made possible by the support of the Office of Naval Research, Project N6onr-246, T.O. IV, NR 173-348.  
<sup>1</sup> Clark, K. E., and Swanson, C. E. Neighborhood reaction to public low-rent housing. *J. appl. Psychol.*, 1951, 35, 342-347.

Table 1  
Opinions Toward Low Rent Housing Project in 1950 and in 1952

	Favor		Oppose		No Opinion or Qualified	
	1950	1952	1950	1952	1950	1952
Total Group: Number	73	159	71	108	44	84
Per Cent	39	45	38	31	23	24
By Intensity of Feeling:						
Very strongly	42%	46%	49%	53%	0%	1%
Rather strongly	33	35	39	33	0	4
Not strongly at all	24	17	12	13	39	21
No answer	1	2	0	1	61	74
Total	100%	100%	100%	100%	100%	100%

expected during this period. These two distributions are presented in Table 2.

That our estimate of the average increment was not too far off is indicated by the similarity of the two distributions.

Table 3 shows that in 1950 a slightly larger proportion of persons with high incomes

tended to favor the project than did those with lower incomes. This was also true in 1952. The 1952 opinions are generally more favorable for all three groups; the lowest income group, however, has a much larger percentage of undecided respondents in 1952 than it had in 1950. Why this should be is

Table 2  
Reported Incomes of Respondents in 1950 and in 1952

1950	1952	N		Per Cent	
		1950	1952	1950	1952
\$5,000 up	\$5,500 up	62	108	33	31
4,000-4,999	4,500-5,499	38	54	20	16
3,000-3,999	3,500-4,499	36	71	19	20
2,000-2,999	2,500-3,499	30	53	16	15
1,000-1,999	1,500-2,499	13	21	7	6
0-999	0-1,499	2	15	1	4
No answer		7	29	4	8
Total		188	351	100%	100%

Table 3  
Opinion on Housing According to Income Level in 1950 and in 1952

Income Level		N		Favor		Oppose		Qualified or No Opinion	
1950	1952	1950	1952	1950	1952	1950	1952	1950	1952
\$5,000 and up	\$5,500 and up	62	108	44%	47%	35%	34%	21%	19%
\$3,000 to 4,999	\$3,500 to 5,499	74	125	39	52	35	28	26	20
Less than \$3,000	Less than \$3,500	45	89	35	40	43	27	22	33

Table 4  
Do You Think Property Values Will Go Up, Down, or Stay the Same?

	N		Go Up		Go Down		Stay Same		Other	
	1950	1952	1950	1952	1950	1952	1950	1952	1950	1952
Total Group	188	351	5%	4%	35%	28%	50%	60%	10%	8%
Favor Project	73	159	8	5	7	11	77	77	8	7
Oppose Project	71	108	3	2	75	60	16	30	7	8
No Opinion or Qualified Opinion on Project	44	84	2	2	18	19	61	68	19	11

not clear. It would seem that these persons became more uncertain about the project as its reality increased.

Tables 4, 5, and 6 present responses to three question requiring prediction about the effects of the project, divided according to responses favoring or opposing the project. These results are particularly interesting since they indicate a lessening of predictions of an unpleasant sort by those who oppose the project. Thus, those persons who *still* say they oppose the project do so with less associated feelings of unpleasant consequences of the project.

A series of information questions was included in the original survey to determine the degree to which persons had become acquainted with specific portions of the plan for the development. Responses in 1950 and 1952 are compared in Table 7. Item numbers refer to the following questions:

1. "About how many families do you understand will be housed in this development?" Correct answer in 1950 was, "120"; in 1952, "184."
2. "How much a month will be charged for

Table 5  
Do You Think This Unit will Bring Undesirable People into Neighborhood?

	N		Yes		No		Other	
	1950	1952	1950	1952	1950	1952	1950	1952
Total Group	188	351	41%	28%	38%	45%	21%	27%
Favor Project	73	159	15	11	73	67	12	22
Oppose Project	71	108	76	58	10	18	14	24
No Opinion or Qualified Opinion on Project	44	84	30	23	25	39	45	38

Table 6  
Will Construction of Development Have Effect on Your Long Term Plans to Stay or Move Out of Neighborhood?

	N		Yes		No		Other	
	1950	1952	1950	1952	1950	1952	1950	1952
Total Group	188	351	28%	12%	64%	77%	8%	11%
Favor Project	73	159	11	1	86	96	3	3
Oppose Project	71	108	55	33	28	47	17	20
No Opinion or Qualified Opinion on Project	44	84	11	6	84	79	5	15

Table 7  
Correct Responses to Information Questions in 1950 and 1952

	N		Question 1		Question 2		Question 3		Question 4		Question 5	
	1950	1952	1950	1952	1950	1952	1950	1952	1950	1952	1950	1952
Total Group	188	351	30%	23%	12%	23%	15%	79%	27%	34%	24%	39%
Favor Project	73	159	32	21	12	25	25	84	30	32	47	52
Oppose Project	71	108	32	29	13	26	13	84	31	41	24	26
No Opinion or Qualified Opinion on Project	44	84	23	19	9	14	2	64	16	29	27	32

rent, heat and utilities for one of these units?" Correct answer in both years was "about \$36 per month."

3. "How many stories do the units have?" Correct answer in both years was, "two."

4. "What is the most money a family can make a year and still rent a place in the development?" Correct answer in both years was, "not to exceed \$2400 plus \$100 per dependent."

5. "If an undesirable family gets into this development will the housing authority be able to get them out?" Correct answer in both years was, "yes."

The percentages reported in Table 7 are the percentages of correct answers. Only question 1 shows a decrease in correct information, perhaps due to the change in correct answer from 120 families in 1950 to 184 families in 1952. The only item on which opponents of the project show less information than proponents is question 5. Question 3 shows a remarkable improvement in information, although it is rather surprising that even though this large and prominent project exists in their immediate neighborhood, 21% of the residents do not know that the buildings are two stories in height!

#### Results from Matched Respondents

The preceding analysis is of interest in describing the total change in sentiment in the community, but yields little information about what happens to the individual respondent as a result of watching this project develop from the planning to the construction stage. Accordingly, from the sample of addresses used in both the 1950 and 1952

surveys, respondents were matched, as nearly as possible, using information on apparent age, sex, and reported education. A total of 171 of the original 188 households were included in the 1952 survey. Of these, 14 were newcomers; i.e., they reported that they were not living in their present house in June of 1950. Of the remaining 157, 66 persons were found with matching characteristics, and are assumed to have been interviewed in 1950 and 1952. Their responses for these two surveys are shown in Table 8.

These results are not in accord with those for Table 1, where the percentage of qualified and no opinion responses remained as high in 1952 as in 1950, and where, apparently, gains in favor were made at the expense of the "oppose" group. These results, for matched respondents, indicate that gains in favor are made at the expense of the qualified or no opinion groups.

Further information on this point is obtained by comparing the responses of matched

Table 8  
Comparison of Responses of Same Persons  
in 1950 and 1952

1950	1952			Total
	Oppose	Qualified or No Opinion	Favor	
Favor				
Qualified or	1	1	21	23
No Opinion	5	9	8	22
Oppose	15	3	3	21
Total	21	13	32	66

households, rather than matched individuals. These results are shown in Table 9. (Note that the following values include the persons already reported in Table 8.)

The results in Table 9 have somewhat more meaning than the preceding ones, since they are based on larger N's, but have lost somewhat in their significance since they represent matched households rather than matched individuals. If we may overlook the latter factor, it seems clear that the most significant changes in response have occurred in the original opposition group. This group shows almost as much shift in opinion as the original no-opinion group, and shifts almost as much to favor as it does to no-opinion.

Table 9  
Comparison of Responses of Same Households  
in 1950 and 1952

1950	1952			Total
	Oppose	Qualified or No Opinion	Favor	
Favor	8	9	38	55
Qualified or No Opinion	12	16	15	43
Oppose	32	15	12	59
Total	52	40	65	157

The total number of households shifting from one position to another, shown in Table 9, is larger than one might have predicted, especially when the direction of shift varies as much as it apparently does. Is it possible that some of the original responses were held with little intensity, and so were almost the same as no-opinion responses? Some evidence on this point is available in the matched-person sample in Table 8 from the response to the question on intensity with which the opinion was held. One might expect the distribution of responses in 1950 of "changers" to be somewhat different from that of the "non-changers." But such is not the case. Although in this group the N's are very small (only 8 persons with an original response of favor or oppose in the matched

sample changed their responses in 1952), the distributions on intensity are still so nearly identical as to suggest that this is not a likely explanation. What seems more likely is that the large number of "changers" occurs only when we interview different persons in the same household. This suggests "division of opinion" *within* households as well as *between* households.

The findings of a study of this sort have considerable significance for those persons associated with planning civic programs, for, in working on plans, one must not only consider the opinions of one's public at the time plans for change are announced, but also the eventual degree of acceptance or non-acceptance of the completed project. In this particular instance the effect of the actual construction of a low-rent housing project was to reduce, but only to a slight degree, the opposition of the neighboring residents to the project.

### Summary

Shortly after the approval in 1950 of plans for a low-rent housing project in a metropolitan area, interviews with neighboring householders were conducted to determine their opinions about this project. Again in 1952, shortly after completion of the construction of the project, but before the new occupants began to move in, interviews were conducted with the original sample of households, and with an additional sample of about the same size. It was found that:

1. About equal numbers favored and opposed the project in 1950, with about one in four undecided. In 1952, about the same proportion continued to be undecided, but the proportion favoring the project had increased slightly (from 39 per cent to 45 per cent). The large number of respondents who continue to be "undecided" in 1952 occurs in spite of the prominence of the project, the considerable publicity given to it, and the controversy about it.

2. Comparison of the responses of persons and households who appeared in both the 1950 and 1952 samples does not indicate very clearly the source of the increased response

in favor of the project, but does suggest that when changes occurred, they were more likely to be from oppose to undecided, or undecided to favor, rather than from oppose to favor.

3. What changes did occur must be considered to be the result of the appearance of the project rather than the characteristics of the new residents, since interviewing was completed before the units were occupied.

4. Of incidental interest is that, by means of several call-backs plus the use of well trained interviewers, the number of refusals to be interviewed was kept below two per cent in both 1950 and 1952, and the number of interviews completed in the two fixed-address samples was maintained at about 96 per cent.

*Received August 11, 1952.*

## Group Performance in a Manual Dexterity Task

Andrew L. Comrey

*The University of California at Los Angeles*

The research to be reported in this article was designed to provide some information on the following questions: (1) how well might the performance of a pair of individuals on a manual dexterity task be predicted from a knowledge of the individual manual dexterity scores of the two persons making up the group; and (2) does the relative level of group performance seem to be more closely associated with the lower or the higher of the individual performers.

### Experimental Procedure

Each "group" studied in this research consisted of two men. One hundred and thirty volunteers were recruited about equally from undergraduate and graduate students to make up 65 groups. No attempt was made to control the placement of individuals into groups. Some pairs were composed of friends, but in the majority of cases the individuals were either unknown to each other or were only casually acquainted.

The subjects in each pair were brought into a well lighted and ventilated experimental room and seated across from each other at a table of approximately office-desk proportions. The experimenter was seated at the far end of this table. In front of each subject was a Purdue Pegboard Test, the two boards touching each other at the ends containing the peg cups. The standard instructions for the Purdue Pegboard, Assembly Task were read to the subjects. Following this, six standard trials were taken, the "Assembly" scores being recorded after each trial for each subject individually. Each subject was able to see how well the other person was doing in comparison with his own performance.

Following the completion of six trials of individual performance, one of the pegboards was removed and the other was placed between the two subjects, the long direction of the board perpendicular to the axis through the subjects, and the cups away from the experimenter. The following instructions were read by the experimenter:

"In the second part of the experiment, you will work on the same type of task except that you will work together rather than individually. First (*subject A—on E's left*) will pick

up a peg and place it in the first hole of the row nearest you. Then (*subject B—on E's right*) will pick up a washer and place it over the pin. Then (*subject A*) will pick up a collar and place it over the washer. Then (*subject B*) will pick up a washer and place it over the collar, completing the first assembly. At the same time, (*subject B*) will pick up a peg with the other hand and place it in the second hole of the row nearest him. Then (*subject A*) will place on a washer, (*subject B*) will put on the collar, and finally (*subject A*) will place on the final washer, picking up a peg at the same time with the other hand and placing it in the hole diagonally across from the assembly being completed. Thus, the assemblies zigzag down the board, each person's assignment alternating on each successive assembly. Now do a few assemblies for practice."

When it was clear that the subjects understood the nature of the group task, six trials of one minute each were taken, scores being recorded for each trial. Scoring was the same as for the individual assembly task. This completed the experimental session, usually taking 25 to 30 minutes. In reading the instructions for the group task, the subject's name was inserted in the appropriate space, italicized in the text given above.

### Treatment of the Data

The statistical analysis of the data proceeded in the following steps:

(1) For each person, individual assembly scores on trials three and five were added together and scores on trials four and six were added together. These "split-half" scores were used to obtain reliability estimates and also were added to give a total individual performance score. The same procedure was followed for the "group" scores.

(2) The members of each pair were designated as "high" or "low," respectively, on the basis of their total individual performance scores computed in (1) above. The person of the pair with the higher total was automatically classified as "high" and his partner was classified as "low." Many individuals in the "low" classification had higher scores than some persons in the "high" classification. This apparently arbitrary method of dividing subjects was adopted because one objective of the experiment was to determine whether the lower or the higher of the two individual performers of a pair would have a greater influence on their group effort. Common sense

Table 1  
Summary of Results

Score	M	$\sigma$	$r_{11}$	Corrected $r$ with			Beta Weight
				High	Low	Group	
High	186	16.5	.90	1.00	.52	.56	.35
Low	173	16.8	.92	.52	1.00	.59	.41
Group	178	19.2	.87	.56	.59	1.00	
$R = .66$				$R^2 = .44$			

might suggest that the pair could do no better than the poorer man, in a normative sense.

(3) Pearson product-moment correlations were computed between the "split-half" scores in the "high" and "low" categories and also for the "group" performances. Thus, the 65 persons in the "low" category each had two half scores. These scores were correlated and the correlation corrected for doubled length by the Spearman-Brown prophecy formula to obtain a reliability estimate for total individual performance scores in the "low" category. The same procedure was followed for those persons in the "high" classification and for the pairs of individuals involved in the "group" performance.

(4) Pearson correlations were computed between "high" and "low" individual performances, between "high" individual and "group" performances, and between "low" individual and "group" performances. These correlation coefficients were corrected for attenuation in both variables involved, using the estimates of reliability obtained in (3) above. The correlations so obtained were treated as estimates of the values which might be expected between the given variables had the measures involved been entirely free of errors of measurement.

(5) A coefficient of multiple correlation between "group" performance and the respective "high" and "low" individual performances was computed, using the coefficients of correlation corrected for attenuation, as obtained in (4) above. The multiple correlation was computed using correlations corrected for attenuation because it was desired to have an estimate of the maximum amount of variance which could be predicted under ideal conditions, i.e., with errors of measurement absent. The idea was to gain some indication of how much variance might be attributable to certain additional unknown variables.

(6) Beta weights for the "high" and "low" scores, respectively, were computed for the regression equation to predict "group" performance scores.

### Results

The results of the statistical analysis have been summarized in Table 1. Inspection of

the scatter plots revealed no indication of curvilinear regressions, although the plot between "high" and "low" scores had a restriction due to the fact that the "low" score of a pair could not be greater than the "high" score.<sup>1</sup> In the first column of Table 1 are listed the total score categories, "high," "low," and "group," standing, respectively, for those total performances as described in the previous section. The means and standard deviations of the three sets of scores are given in the second and third columns, respectively. These are based on the totals of the last four of six trials. This procedure was decided upon in advance to obtain more stabilized results. In the fourth column are given the reliability estimates as obtained by the procedure described in (3) of the last section. The next three columns of Table 1 give the intercorrelations of the total score variables, corrected for attenuation. The steps were described in (4) of the section on treatment of the data. The last column contains the beta weights for predicting "group" performance from "high" and "low" individual performances. The multiple correlation,  $R$ , and  $R^2$ , as described in (5) of the last section, are given in the bottom row of the table.

### Discussion

Information pertaining to the first question in this research is given by the multiple correlation coefficient between "high" and "low" scores and the "group" score. The square of that coefficient indicates that 44 per cent of

<sup>1</sup> To determine the effect of this artificial restriction, 65 pairs of two-digit numbers were taken from a table of random numbers, placing arbitrarily the higher of the two numbers in the first group and the lower of the two in the second group. The resulting Pearson correlation was .56.

the variance in an errorless measure of group performance could be predicted from a linear combination of perfectly reliable "high" and "low" individual scores. The percentage which can be predicted by fallible measures would be less.

Three possible explanations of this result will be given. First, the task itself may actually be significantly different in its nature from the individual assembly task. It was necessary for the subjects to alternate operations on succeeding assemblies when working together which was not the case in individual operations. In future work, a redesigned individual task will be used which requires the subject to interchange the sequence of hand movements on alternate assemblies. This should make the operations in the individual task more like those in the group task.

Even though the previously mentioned difference in the individual and "group" tasks were eliminated, this would by no means indicate that the "group" task would then be the same to the participating persons as their individual tasks. The two tasks are probably different to each individual not only because of an intrinsic difference in the sequence or character of the operations but also because the group situation brings in new elements requiring the utilization of different abilities. The person must anticipate the moves of his partner to achieve a smooth performance. In short, it is suggested that there may be a group of abilities possessed to different degrees by different individuals which determine in part how well they will perform in certain group situations. These new abilities may be independent of those which determine the performance of the same operations in the same sequence by the individual as a single performer.

A third possible explanation of these results lies in the hypothesis of interactions among individuals. It may be that some or even all subjects will work more effectively with some individuals than with others. Under this hypothesis, variations in group performance may be substantially influenced by the extent to which persons are paired who will work best together.

The second objective of the experiment was to determine if the lower individual performer of a pair influenced the group performance more than the higher individual performer. This question may be answered by an examination of the correlations of the "low" and "high" scores with the "group" scores. The correlations, corrected for attenuation, were .53 and .50, respectively. The difference in effect on group performance by the "high" and "low" pair members is so slight as to be of no practical consequence. Thus, group performance here seems to be a function of the average of the two individual scores. Under these conditions, for a given group of workers, there would seem to be little to gain by trying to pair off the high ones and the low ones, expecting thereby to get more over-all production from the group as a whole. This conclusion naturally presumes a similar type of prediction situation and the lack of further information beyond that which was available here.

The same statistical treatment was also given to the data from the first two trials. Reliabilities were somewhat lower, .76, .82, and .74 for the "high," "low," and "group" scores, respectively. Intercorrelations among these scores, corrected for attenuation as before, were "low-high," .53, "low-group," .67, and "high-group," .58.  $R^2$  was .51. Thus, during the practice trials, the "low" men influenced the group performance slightly more than they did in later trials. Also, during these trials, the group scores were more highly related to the individual performance scores than during the test trials, as shown by the higher  $R^2$ . The stabilized performance results probably have the greater practical value, however.

#### Summary

Sixty-five pairs of volunteer male university students were given six trials on the Purdue Pegboard, Assembly Task, and six trials on the Assembly Task with the two members of each pair working together on the same assemblies rather than individually on separate boards. The members of each pair were divided on the basis of the total of the last four individual trials, Assembly Task, into "high"

and "low" categories. Reliabilities were determined for "high," "low," and "group" performances, using alternate trials and correcting for doubled length. Correlations of the "high" and "low" performances with the "group" performance and with each other were computed and corrected for attenuation. The multiple correlation and regression weights were obtained for predicting "group" performance from "high" and "low" individual performances.

The results showed that less than half the group performance variance could be predicted from a knowledge of the individual performances, even with the effect of errors removed. It is suggested that manifest differences between the "individual" and "group" tasks, interactions among individuals, and a constellation of abilities in the general area of cooperation may account for the variance not predicted by perfectly reliable individual performance scores.

The level of group performance was only slightly more dependent on the "low" individual performances. For all practical purposes equal weights could be used for "high" and "low" scores in predicting "group" performance.

Two practical implications of the results of this experiment are as follows. First, in industrial situations where two or more individuals must cooperate on a given task, it must not be assumed that individual performances on a similar type of task will account for most of the variation in group performance. Secondly, for a given group of persons there seems to be little point in taking the trouble to pair them on individual ability in a related type of individual task since group performance seems to be dependent on the approximate average of their individual scores rather than the high or low individual performance.

*Received August 4, 1952.*

## Response Time as an Indicator of Color Deficiency \*

Sherman Ross and 1st Lt. John L. Fletcher, MSC, USA

University of Maryland

In 1907 Froeberg (2) reported that reaction time varied inversely with the intensity of the stimulus. Since that time reaction time measures have been put to a variety of uses. Steinman (6) reported that simple reaction time to stimulus change was an adequate method for studying sensitivity to stimuli. She also found that reaction time decreased as the magnitude of the change increased. In a study whose purpose was to determine the speed and accuracy of discriminations of hue, brilliance, area, and shape of visual stimuli, J. B. Reed (4) found that as the difference between two areas or hues is increased, discrimination time is decreased.

When pseudo-isochromatic plates are used, the subject reacts to a stimulus complex. This suggests that if the discrimination required is difficult, the response time would be longer than if the discrimination were easy. Further, if color contrast is absent on the test plate the subject would hesitate and seek other cues, such as differential brightness, as a basis for responding.

Reports from several investigators lend support to this notion. J. D. Reed (5) studied reactions to a complex submarine signal panel board, and reported that use of reaction time measures revealed the increased difficulty of discriminations for color defectives. Also studies by Pickford (3) and Sultzman (7) refer to hesitancy on the part of the color defective individuals.

On the basis of these reported observations of hesitancy on the part of color defective individuals, the following hypothesis was tested: color "defective" individuals will have longer

mean plate response times than individuals classified as "normal."

### Method and Procedure

**Test.** The color test used in this experiment consisted of a set of 15 pseudo-isochromatic plates (14 diagnostic and 1 demonstration) selected from the American Optical Company test<sup>1</sup> by Farnsworth and called the "Proposed Armed Forces Color Vision Test for Screening" (1).

**Subjects.** A total of 136 students (108 male, 28 female) from the University of Maryland were used in the experiment.

**Method.** Test plates were presented singly to the subjects in the order prescribed by Farnsworth (1). The subjects were tested twice in succession. Half of the subjects were given a *criterion* trial first, while the remaining half were given a *test* trial first. The criterion trial was conducted exactly as recommended by Farnsworth (1) except that the subjects were instructed to respond to the plate as soon as possible. If no response was made in 3 sec., the plate was removed. The test trial differed from the criterion trial only in one respect, i.e., response times were taken for each plate in the test trial.

**Apparatus.** Illumination was provided by a Macbeth Daylight lamp (No. ADE 10) as suggested by Farnsworth (1). The test plates appeared against a flat black background and were placed on a bracket which slid rapidly into a viewing aperture. Presentation of the plate started an electric chronoscope calibrated in .01 sec. The subject's verbal response activated a voice key and stopped the timer. Verbal reports and response times as described above were recorded.

### Results and Discussion

The error scores made by the subjects on the criterion trial were used to classify the

\* The writers would like to express their sincere thanks to Lt. Comdr. Dean Farnsworth, M.S.C., USNR, Head, Visual Engineering Section, U. S. Naval Medical Research Laboratory, New London, Conn., for his review of the manuscript. The opinions or assertions expressed in this paper are those of the writers and are not necessarily those of the military departments.

<sup>1</sup> Pseudo-Isochromatic Plates for Testing Color Perception (Revised Selection, American Optical Company).

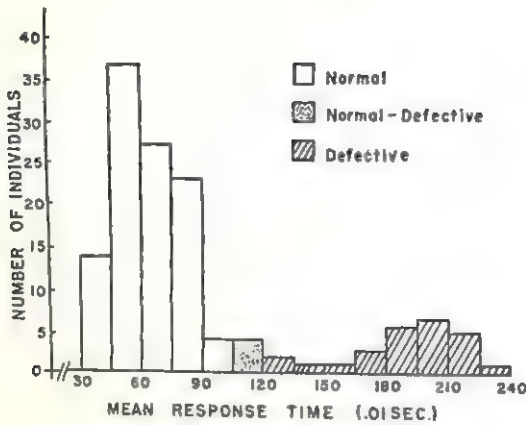


FIG. 1. Mean plate response time frequency distribution for normal and color defective subjects ( $N = 136$ ).

subjects as normal or defective. The norms recommended by Farnsworth (1) were used in the classification; i.e., four errors or less for normals, five errors or more for defectives.

Using this criterion, 110 subjects (84 male and 26 female) were classified as normal. The mean plate response time for the normal group was .66 sec., ranging from .31 sec. to 1.42 sec. The defective group had 26 members (24 male and 2 female). Mean plate response time for the defective group was 1.93 sec., ranging from 1.24 to 2.40 sec. Fig. 1 shows the mean plate response time distribution for all subjects, normal and defective. It should be noted that subjects classified as normal by the criterion test are also categorized as normal by mean plate response times, and that the deficient subjects, with the exception of two cases, are classified as defective by both error score and response time measures.

The  $t$ -test was used to test the significance of the difference between the means of the distributions of mean plate response times for the normal and defective subjects. The difference was found to be significant at less than the .001 level of confidence.

Table 1 presents a plate-by-plate analysis of the errors made by the normal and defective subjects. The first analysis of the error scores consisted of a  $t$ -test to determine whether or not a significant practice effect existed. We were unable to reject the null

hypothesis for the defective group, but were able to reject the null hypothesis at less than the .01 level of confidence for the normal group. This could mean that there is no appreciable change in error scores for defective subjects. On the other hand, normal subjects made significantly fewer errors on their second test trial.

It may be noted from Table 1 that normal subjects in varying numbers missed plates 2, 6, 10, 11, 14, and 15. Specifically, 51 per cent of the normal subjects missed plate 6, and 39 per cent missed plate 15 on the first trial. Errors were made on every plate except the demonstration plate by the defective

Table 1

Number of Errors per Plate by Normal and Defective Subjects for the First and Second Trials

Plate No.	Normal (N = 110)		Defective (N = 26)	
	Trial 1	Trial 2	Trial 1	Trial 2
2	9	3	20	22
3			25	25
4			16	17
5			20	23
6	56	38	25	24
7			20	17
8			16	16
9			18	7
10	2	2	16	16
11	14	14	24	24
12			19	20
13			13	14
14	1	1	20	18
15	43	23	25	25

individuals. Plates 2, 3, 5, 6, 7, 11, 14, and 15, however, were missed by a larger number of these subjects than the other plates.

It was stated earlier that response time measures might prove useful as an indicator of color deficiency. An examination of Fig. 1 reveals that the mean plate response time frequency distribution of the normal and defective subjects overlaps in only one interval, 1.20-1.35 sec. With the exception of the two cases represented, all subjects with a mean plate response time of 1.50 sec. or over (or an over-all response time of 21.0 sec.) could

be classified as deficient, and those with response times less than 1.50 sec. as normal.

It may well be possible that the limit of 3 sec. employed put an artificial limit on the response times. A color-blind individual would not necessarily, at the end of this period make the correct response or any response. On the other hand, a normal or near-normal subject, who usually eventually makes correct responses, might well be categorized by a response time method. For further investigation of the practical usefulness of this method, the 3-sec. ceiling should be extended.

The problem in classification is that of the near-normal subjects. The number of borderline subjects in the present experiment was limited, and these borderline subjects constitute the difficult classification problem. Most tests will roughly separate the extremes of the population. The problem is to devise tests to select the Class II group (mildly defective color vision) as classified by Farnsworth (1).

Further research is indicated along this general line since it is entirely possible that response time measures could serve in classification. It is possible that memorization of the plates could be detected by this method. Response time measures could readily be used in military and industrial situations.

### Summary

Response time measures to 15 selected plates of the AO pseudo-isochromatic test for color perception were secured for 136 college students (28 females, 108 males). Of this group, 110 were classed as normal, 26 as de-

fective, on the basis of their error scores. Each subject was given two successive tests: criterion and test trials. Mean plate response times between the normal (0.66 sec.) and defective (1.93 sec.) groups were found to differ significantly. Practice effects were noted within the normal group, but were not found in the defective group. It was concluded that response time measures could be used in the separation of color normal from color defective individuals.

Received August 14, 1952.

### References

1. Farnsworth, D. *Proposed armed forces color vision test for screening*. U.S.N. Submarine Base, Medical Research Laboratory, Report No. 180, 1951, 10, 146-155. (Color Vision Report No. 24.)
2. Froeberg, S. The relation between the magnitude of the stimulus and the time of reaction. *Arch. Phil., Psychol., Sci. Meth.*, 1907, No. 8, 38.
3. Pickford, R. *Individual differences in colour vision*. London: Routledge and Kegan Paul, 1950. Pp. 386.
4. Reed, J. B. *The speed and accuracy of discrimination differences in hue, brilliance, area, and shape*. U.S.N. Special Devices Center, Port Washington, L. I., N. Y. Tech. Rept. SDC-131-1-2, Sept. 1951.
5. Reed, J. D. A note on reaction time as a test of color discrimination. *J. exp. Psychol.*, 1949, 39, 118-121.
6. Steinman, A. Reaction time to change. *Arch. Psychol.*, 1944, 41, No. 292. Pp. 60.
7. Sultzman, J. H., Lieut. (MC) USNR. *Comparison and evaluation of the American Optical Co. Pseudo-Isochromatic Plates, First and Second Editions*. U.S.N. Submarine Base, Medical Research Department, BuMed-X-480 (Au-255-p), 16 July 1945.

## The Effect of Set on Performance in a "Trouble Shooting" Situation

Nicholas A. Fattu and E. Victor Mech<sup>1</sup>

*Institute of Educational Research, Indiana University*

With equipment-systems periodically gaining in complexity, both industry and the Armed Forces are faced with the task of training personnel in the maintenance of these equipment-systems. However, the problem of how to systematically train maintenance men has thus far received meager experimental treatment.

It has frequently been postulated that the trouble shooting process of locating defects in equipment-systems is similar to that encountered in problem solving situations. When a mechanic is confronted with an equipment-system that is "malfunctioning" and no appropriate response is available, the process of locating the defective part clearly possesses the properties of a problem situation. Unfortunately, the gap between the problem solving literature and understanding complex trouble shooting processes is not easily bridged. Although several promising concepts are embodied in the problem solving literature apparently there is no simple transplanting of these notions to problems encountered in the trouble shooting of complex equipment-systems.

The purpose of the present experiment was to test a concept found important in previous problem solving studies by applying it to a more realistic problem situation.

The studies of Maier (2, 3) suggest that a knowledge of the required parts of a solution does not necessarily mean the occurrence of a solution. Evidence is presented that indicates additional information in the form of a set is necessary; necessary in the sense that the additional set clears the way and increases the probability of a correct solution.

The hypothesis tested in this experiment was that the ability to "trouble shoot" or locate defects in an equipment-system entails

more than being trained in the basic components, or essential parts of the equipment.

It was postulated that in addition to teaching basic components something more was required in the form of a set of principles dealing with systematic location of a "malfunctioning" component.

### Method

**Apparatus.** The apparatus in Figure 1 is called a gear-train consisting simply of a set of gears and shafts mounted on a piece of aluminum  $\frac{1}{2}$  inch thick, 29 inches in length, and 20 inches in width. The gear-trains were arranged to form two series and four parallel channels that provided for crossed information chains.

Two operating controls, A and B, provided the input necessary to obtain the desired motion. The motion was transferred through the gear-trains and as an end result closed a switch that caused a series of red lights to illuminate the control panel.

The red lights would illuminate only if the equipment was functioning properly and control A was turned 13 times and control B, 12 times. When the appropriate number of turns was made and the expected end result (control panel lighting up) did not occur, this indicated to S that there was a "malfunction" in the gear-train.

**Malfunctions.** For the purposes of the experiment only one class of "malfunction" was utilized. It was a defect of the "slipping gear type" produced by loosening a set screw and found by the authors in a previous study (1) to be fairly difficult. Each S received six malfunctions of this type on the pre-test and six malfunctions of the same type on the post-test, making 12 malfunctions that each S was required to locate.

Ss were presented the malfunctions in a random order, and, in addition, each of the malfunctions was inserted at a location determined from a table of random numbers.

**Procedure.** Ss were run individually, and immediately upon entering the laboratory for the first time E gave S the Standard Operating Procedure (hereafter referred to as S. O. P.) for the apparatus. The S. O. P. consisted of turning control handle A, 13 turns and control handle B, 12 turns, and if the red lights on the control panel did not light up, it indicated to S that something was wrong with the gear-train and the task was to "trouble shoot" the equipment. After the S.

<sup>1</sup> The authors are indebted to Mr. Walter Ciszczon for material contributions to apparatus, and to Mr. Jasper Smaliks for aiding in the derivation of the symptom analysis "trouble shooting" method.

O. P. orientation, *S* was given six malfunctions or problems to locate. Each problem was given singly, *S* being placed in an adjoining room while the malfunction was being inserted by *E*. A time limit of 15 minutes was allowed for each malfunction.

When each *S* had completed working on the initial six "malfunctions," the following procedure was followed. *Ss* in Group 1 were taken to an adjoining room for a 20-minute period during which they were allowed to read a current Life magazine. *Ss* in Group 2 received a tape-recorded "basic knowledge" lecture. Integrated with the lecture were slides projected on a screen with a 35 mm. camera. The rationale of the lecture was to convey the basic nomenclature and function of the gear-train apparatus. Included were such concepts as transfer of motion, and the function of gears, bearings, and shafts. Group 3 received the basic knowledge information given to Group 2, and, in addition, was given a tape-recorded lecture on how to "trouble shoot" the gear-train. This "trouble shooting" lecture was based on a simplified version of a previously developed symptom analysis guide to "trouble shooting." The lecture stressed starting with the greatest magnitude of error, locating the first correctly operating component nearest the greatest error, and once these two points were bracketed, locating the defect somewhere between.<sup>2</sup>

<sup>2</sup> The symptom analysis "trouble shooting" lecture has been filed with the American Documentation Institute. Order Document 3968 from ADI Auxiliary Publications Project, Photoduplication Service, Li-

*Subjects.* The *Ss* were 54 college students enrolled in the School of Education at Indiana University. Three groups of 18 *Ss* each were used in the experiment. Assignment of the 54 *Ss* to each of the three groups was done from a table of random numbers. Participation in the experiment was required in order to eliminate the bias often found by asking for volunteers.

## Results

The raw data for this experiment are the post-test gains or the number of malfunctions correctly located minus the number located on the pre-test, and the total time required to reach a decision as to the location of the malfunctions. Arriving at a decision, however, does not necessarily mean that it was a correct one.

Figure 2 illustrates graphically the post-test gains made by the three groups of subjects in "trouble shooting" the gear-train apparatus. Statistical significance of post-test gains was tested by an analysis of variance of the thirteen possible scores ranging from 6 to - 6.

brary of Congress, Washington, D. C., remitting \$1.25 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.25 for photoprints.



FIG. 1. The gear-train apparatus.

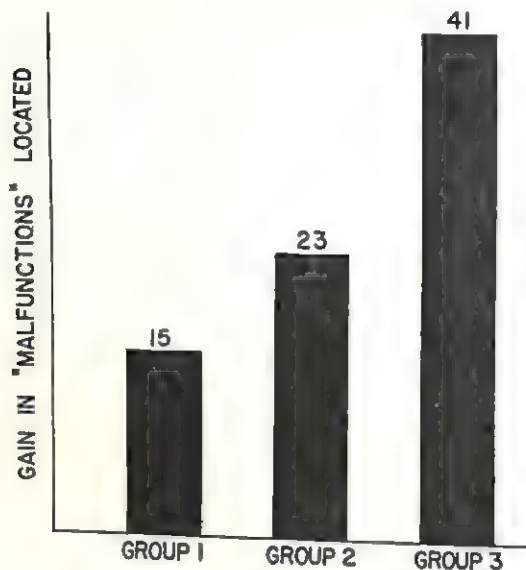


FIG. 2. Post-test gains in malfunctions correctly located over the initial measures.

Prior to computing the analysis of variance, Bartlett's  $\chi^2$  test was utilized to test the homogeneity of variance of defects located. Since the  $\chi^2$  value of 2.26 did not reach the 5% level ( $df = 2$ ), the hypothesis of no differences among group variances could not be rejected.

The analysis of variance performed on post-test gains is summarized in Table 1. The obtained F value of 9.36 for 2 and 51 degrees of freedom was significant at the 1% level of confidence.

It appears, then, that the significant gains demonstrated by the performance of Group 3 subjects can defensibly be attributed to the effects of the "trouble shooting" lecture that the two remaining groups did not receive.

However, a glance at Figure 3, showing the

Table 1

Analysis of Variance of Differences Between Initial and Test Location of "Malfunctions"

Source of Variance	df	Sum of Squares	Mean Squares	F
Between Groups	2	27.88	13.94	9.36*
Within Groups	51	76.01	1.49	
Total	53	103.89		

\* Significant beyond the 1% level of confidence.

Table 2

Covariance Analysis for Time to Decide the Location of Pre- and Post-Test Malfunctions

Source of Variance	Sum of Squares	df	Mean Squares	F
Total	618.37	52		4.07*
Within Groups	531.79	50	10.64	
Adjusted Means	86.58	2	43.29	

\* Significant at the 5% level of confidence.

time in minutes for the groups to reach a decision with respect to where various malfunctions were located, points out an interesting reversal. Although Group 3 was superior in locating defects under test conditions, it is clear they failed to decrease subsequent "trouble shooting" time, while the time required by the remaining two groups was reduced. In order to test for differences with regard to time taken to reach a decision as to the location of defects, an analysis of covariance, shown in Table 2, was carried out between the pre-test and post-test time measures.

The F of 4.07 was significant at the 5% level indicating that the means of the group on the post-test time measures cannot be accounted for by differences in mean level of

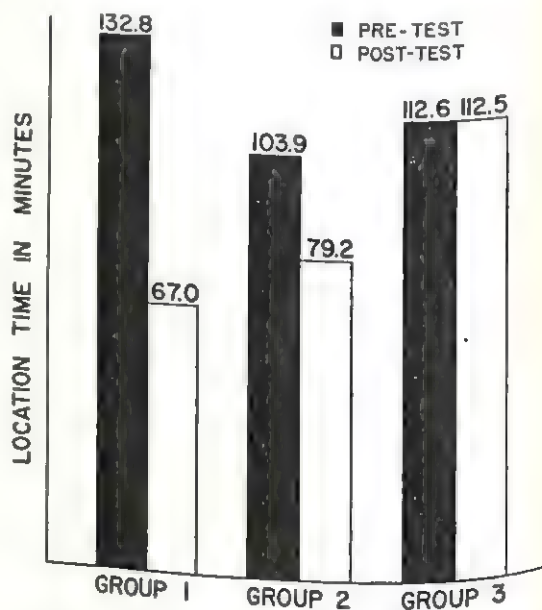


FIG. 3. Comparison between the pre- and post-test of total time required to decide the location of malfunctions.

initial ability as measured in the pre-test trials.

Accordingly, the results of this study are that the additional "trouble shooting" lecture acted to produce a differential effect on subsequent performance in locating gear-train defects. The group which received both the basic knowledge and "trouble shooting" sets did not appreciably reduce the time scores in comparison with the remaining groups.

Time is a rather dubious criterion of performance in the trouble shooting situation. A comparison of Figures 2 and 3 (gains and time) suggests that the longer time required by Group 3 may be attributed to deliberation required for an accurate judgment, while in Group 1 the small time required might be attributed to snap judgment.

In the final analysis, the findings suggest that besides learning about various components of an equipment-system, systematic training in "trouble shooting" methodology is required in order to obtain efficient results.

Intensive research with more complex equipment is needed to determine the additional skills and knowledges required to "trouble shoot" successfully. Such a problem as determining which trouble shooting procedure is more generalizable than another, and testing its transfer power on succeeding complex equipment is, indeed, a fascinating laboratory challenge. By using the laboratory method on these and related problems, much useful information about problem solving in general, and the significant variables of "trouble shooting" in particular could be systematically obtained.

#### Summary

The experiment reported tested the hypothesis that ability to "trouble shoot" or locate defects in a specified equipment-system re-

quires something more than being trained in the parts or components of that equipment-system. Fifty-four undergraduate students enrolled in the School of Education at Indiana University participated in the experiment. Certain training factors were common to the three groups. All Ss received identical indoctrination in the Standard Operating Procedure for a gear-train apparatus, after which each S was given six problems or malfunctions to locate in the equipment. This procedure was used to obtain a pre-test measure of "trouble shooting" ability on the gear-train. After the initial measure, Group 1 received no further information, Group 2 received a tape-recorded basic knowledge lecture that explained the nomenclature and functioning of the gear-train, while Group 3 received the basic knowledge lecture plus symptom analysis lecture designed to aid in "trouble shooting" the gear-train apparatus.

The post-test gains indicate that the additional "trouble shooting" lecture acted to produce a significant gain in malfunctions correctly located. Time required, however, decreased for Groups 1 and 2, but remained constant for Group 3. It is suggested that time is a dubious criterion of "trouble shooting" performance.

Received March 6, 1953.

Early publication.

#### References

1. Fattu, N., and Mech, E. V. Interruption: Its effect upon performance in a trouble-shooting situation. *J. Psychol.*, in press.
2. Maier, N. R. F. Reasoning in humans: I. On direction. *J. comp. Psychol.*, 1930, 10, 115-143.
3. Maier, N. R. F. Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *J. comp. Psychol.*, 1931, 12, 181-194.

## An Evaluation of Two Experimental Charts as Navigational Aids to Jet Pilots \*

John E. Murray

*Dunlap and Associates, Inc., Stamford, Connecticut*

A modern trend in the progress of aviation is toward the provision of improved facilities for the pilots of high-speed, high-altitude aircraft. One of the basic tools required by such pilots is the aeronautical chart. A review of the literature and an examination of current charts show that existing charts fail, in some respects, to provide the pilots of high-speed, high-altitude aircraft with a highly effective navigational tool. Much of the material presented on the charts is superfluous: some of the natural features cannot be seen from high altitudes, and much of the chart content cannot be absorbed in the time available for navigation at high speeds.

An increase in the number of charts was not accompanied by a judicious selection of the chart content. As more aeronautical information became available, it was added to the basic chart without consideration of the flight and navigational requirements of modern planes and air operations.

This procedure resulted in the production of all-purpose charts: charts for use in any aircraft on any type of flight. These charts are cluttered with information, difficult to read and inconvenient to use because of their size. To overcome some of these defects, special purpose charts were designed but without the use of adequate criteria in the selection and presentation of information. Moreover, there seems at present to be no well established methods by which aeronautical charts can be evaluated. Even more striking is the fact that, in the history of chart production, very little systematic study has been made of the pilot's task of interpreting the information presented on the charts. Consequently, the major objective of this study was to devise experimental techniques ap-

plicable to the evaluation of principles of chart construction.

From the point of view of the psychologist, this study is valuable in that it demonstrates the application of psychological methodology to the problem of chart evaluation. From the cartographer's viewpoint, the experimental evaluation of charts yields two types of information which can serve in the future course of chart development: (1) a test of the applicability of general principles and techniques of chart construction; and (2) the relative value of different methods of presenting information.

In order for information to be used most efficiently by the pilot, it should obviously be displayed so as to provide maximum legibility. This involves a determination of the best method of presenting chart information and requires a study of the contributions of color, type of symbol, size and style of printed type, and other related items to legibility.

There are two basic ways in which a chart can be evaluated. One is subjective and depends upon pilots' opinions which can be gathered from interviews and systematic questionnaires; the other is objective and requires the collection of performance test data of various sorts. Both methods have been employed in the present study. The design of this study involves the following steps:

1. The selection of specific features to be included and evaluated on experimental charts.

2. The preparation of tests to measure the readability of charts.

3. The preparation of a test to measure the effectiveness of charts in representing what the pilot sees from the air.

4. The construction of a questionnaire to determine the pilot's attitude toward experimental charts in terms of their content and practicability for actual flight conditions.

The results of the first step are embodied in the structure of two experimental charts.

\* This research was supported under the terms of the contract between the Office of Naval Research, and Dunlap and Associates, Inc., Contract Number N8onr 641-05. This paper is a summary of Report No. 641-05-6 under that contract.

These charts differ in the amount, kind and method of presenting information to the pilot. The significant differences between the two experimental charts are displayed in Table 1. The precise objective of this study was to determine the relative effectiveness of the present World Aeronautical Chart (WAC) and the two experimental charts, the XJN Chart produced by the Aeronautical Chart and Information Service and the XDA Chart designed for the Office of Naval Research. Representative samples of the three charts are presented in Figure 1.

### Evaluation Procedures

**Readability Tests.** Tests were designed to determine the speed and accuracy with which pilots can find and use information contained in the charts. Given the task of reporting certain specified information, the speed and accuracy of performing this task with each

chart can be taken as an index of the effectiveness of their presentation. In effect, these are tests of legibility or ease of reading. This legibility is a function of type size, symbols used, color and the density of the information shown on the chart. The relative superiority of the charts can be determined for those features in which they differ. The following tests of readability were constructed:

**Part I. Airport Information.** Flight lines were drawn connecting fourteen airports. The subject was required to give the airport type, elevation, runway length and available electronic facilities for each of the airports specified. The maximum possible score on this test was 50.

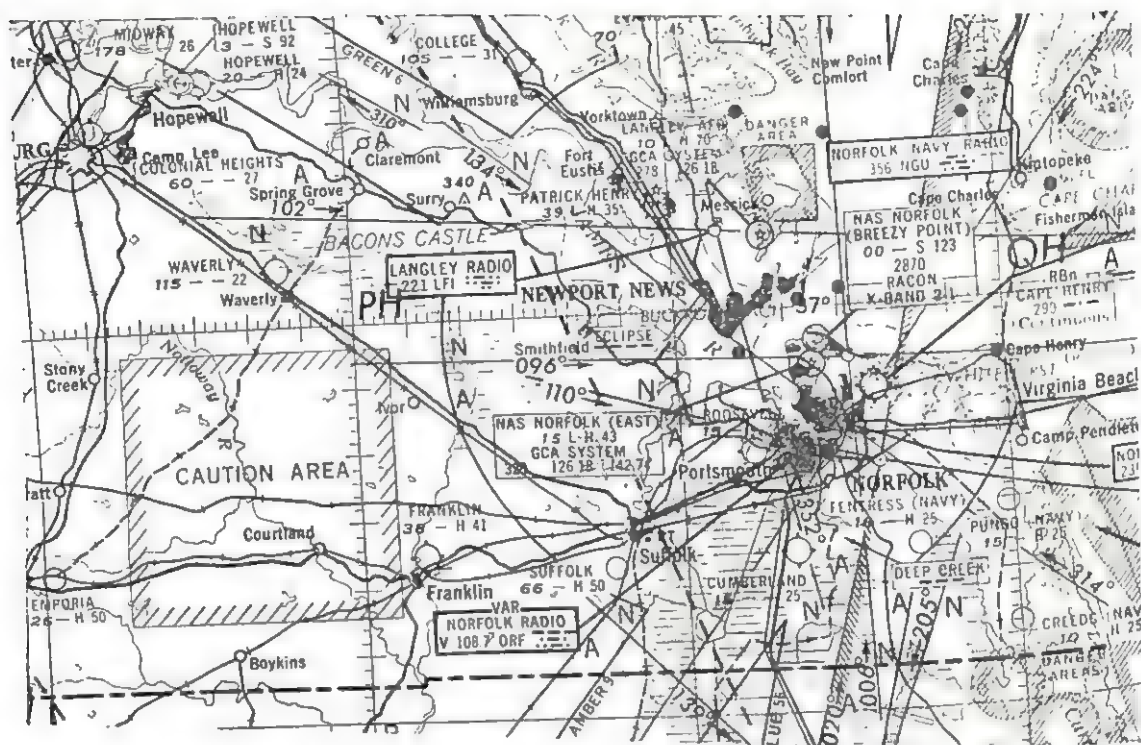
**Part II. Radio Information.** Similar flight lines were drawn connecting various radio aids on each chart and the subject was required to give the type, frequency and call letters for each radio aid specified. The maximum possible score was 42.

Table 1  
Differences in the Presentation of Specific Features on the XJN and XDA Charts

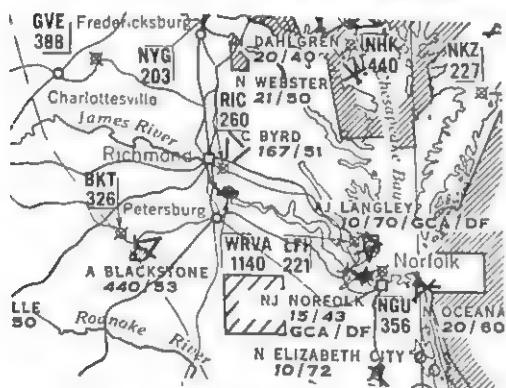
Feature	XJN	XDA
<b>Front of Chart<sup>1</sup></b>		
1. Color of land and water areas	Yellow and blue	Green and blue
2. Terrain features	Hypsometric tints Contour lines Spot elevations Predominant in yellow	Shadient tints to approximate three-dimensional view Highest peak only Subdued in gray
3. Cities	Roads and railroads differentiated	Roads and railroads indicated by same symbol
4. Transportation lines	Jet and military airports shown by runway pattern; civil shown by circle; lighting and surface facilities indicated	All airports shown have adequate lighting and hard surface runway and are represented by runway pattern; type of airport shown in data note; GCA and DF facilities indicated
5. Airports	Shown by dotted lines	Not shown on this chart
6. State names and boundaries	Both airports and radio information in magenta	Airport information in blue; radio information in an improved magenta
7. Colors of symbols	Shown by solid lines	Not presented
8. Navigation light lines	Along edge from 0 to 1,000 miles	Starts from 0 at either end toward 500 in center; in bold type on both sides of chart
9. Distance scale		
<b>Back of Chart<sup>2</sup></b>		
10. Radio beacon	Symbol prominent	Symbol subdued
11. Broadcasting station	Symbol subdued	Symbol prominent
12. Radio range	Shows N quadrant	Differentiates terminal and non-terminal ranges; shows inbound magnetic headings

<sup>1</sup> On the XJN chart, radio and airport information are presented in the same color and the symbols for each differ on the front and back of the chart; on the XDA, airport and radio information are differentiated in color but the symbols for each type of data are consistent on both the front and back of the chart.

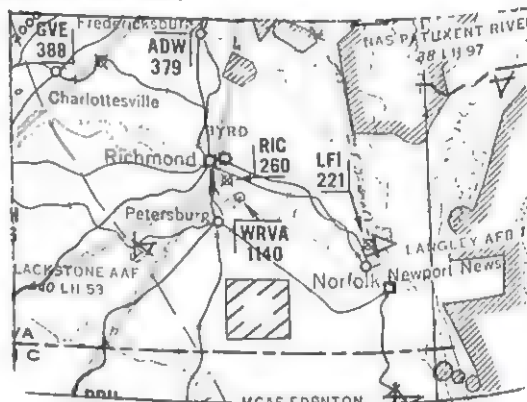
<sup>2</sup> On the back of the chart, XJN shows Morse code, reporting points, fan markers, dumb-bell markers, airways; XDA does not present these features but includes the Atlantic coast line and a list of YG stations.



World Aeronautical Chart (WAC)



XDA Chart



XJN Chart

FIG. 1. Sample sections of the WAC, XJN, and XDA charts.

**Part III. Natural and Cultural Features.** The subject was required to read and interpret various features pertaining to terrain, roads and railroads, cities, rivers, etc., used for navigation in cross-country flight. The maximum possible score was 15.

**Part IV. Aerial Photographs.** This test was designed to measure the individual's ability to read an aerial photograph and to determine its geographic location on the chart. Seven photographs were selected from a series taken at an altitude of approximately 40,000

feet on a flight from Dayton, Ohio to Washington, D. C. The subject was required to locate the area pictured in each photograph on the test chart provided.

Each experimental session was prefaced by an introductory statement covering the purpose of the study and the experimental procedure to be followed. Time limits of five minutes each were imposed on Parts I and II; seven minutes each were allowed on Parts III and IV. Each session required approximately 45 minutes of which 24 minutes were

for working time and the remainder for preliminary instructions.

The tests were administered to groups of 20 to 25 pilots at each session. A total of 72 Navy pilots were tested on the XJN Chart, 66 on the XDA Chart and 60 on the WAC chart for a total of 198 pilots.

*Item Analysis.* As a more refined measure of the effectiveness of the charts, each of the four tests was subjected to an item analysis. The number of individuals who marked each item correctly was determined and a comparison among the three charts was made on each item. In those instances where items were incorrectly marked, the frequency with which other alternatives were chosen was recorded.

*Questionnaire.* To elicit pilot preferences for specific features on the experimental charts, a questionnaire was distributed to another group of Navy pilots. The questionnaire consisted of a series of questions concerning the features which were differently presented on the two charts. In each question, the pilot was asked to state his preference for one of the charts in regard to some specific feature. Where applicable, reasons for the choice or preference were also requested. Free comments, whether favorable or unfavorable, were encouraged as much as possible. In all, 43 pilots were interviewed with the questionnaire either on an individual basis or in small groups of two to four men each.

### Results

*Readability Tests.* The mean score obtained on each test for each chart is presented in Table 2. To determine the effectiveness of each of the charts, the test scores were compared by the standard *t*-test techniques. The test results indicate that airport, radio and cultural information can be read more quickly and accurately on the experimental charts than on the traditional WAC chart. The XDA is significantly superior to the XJN Chart in presenting airport information. This superiority is probably due to the prominence of the airport symbol and the simplicity of the corresponding data note.

Only minor differences exist among the charts when used to identify locations of aerial photographs. This finding would im-

Table 2  
Mean Scores Obtained on Readability Tests  
for the Charts Specified

Chart	No. in Group	Mean	Standard Deviation
Test I			
Airport Information			
WAC	60	37.1	7.31
XDA	66	46.3	6.46
XJN	72	43.0	7.41
Test II			
Radio Information			
WAC	60	34.5	8.76
XDA	65	39.9	5.28
XJN	71	39.0	5.51
Test III			
Cultural Features			
WAC	60	8.3	2.32
XDA	66	12.7	1.99
XJN	72	12.7	1.78
Test IV			
Aerial Photographs			
WAC	60	3.65	1.22
XDA	66	3.23	0.97
XJN	72	3.47	0.76

ply that the reduction in the amount of detail on the experimental charts does not hinder the pilot's identification of reference points.

*Item Analysis.* The data from the item analysis clearly show the relative value of each of the charts as a means of presenting information to the pilot. Economy of space does not permit the inclusion of the data obtained for each item. The important differences among the charts can be summarized as follows:

1. In the time limit allowed, fewer items were completed on the WAC Chart than on either the XDA or XJN Charts. This difference seems to be due to the mass of information shown as well as to the unsystematized placement of the data notes on the WAC Chart. Furthermore, the size and scale of the WAC Chart make it awkward to manipulate and difficult to locate the information required.

2. The runway patterns on the XDA and XJN Charts were more effective than the traditional circular symbols on the WAC Chart.

3. Data notes are more readily identified when placed closely to their related objects. Misidentification of certain airports on the XDA Chart, for example, resulted from improper placement of the data notes pertaining to these airports.

4. Security areas are best shown on the XJN Chart. This seems to be due to the type size and face used in presenting these areas.

5. In presenting terrain features, the XDA Chart is superior to both the XJN and WAC Charts.

*Questionnaire.* The preferences of pilots for the specific features on the two experimental charts are as follows:

1. Printed material on the XDA Chart is more easily read although the chart has a more cluttered appearance.

2. The mileage scale on the XDA Chart is preferred. The scale should range from 0-500 miles from either end of the chart and it should be presented in the same manner on both sides of the chart.

3. Runway patterns on the XDA Chart are preferred by 93 per cent of the pilots interviewed. It is considered desirable to present only those airports with adequate landing facilities for jet aircraft.

4. The bold type for airport information on the XDA Chart is preferred and GCA and DF facilities are highly desirable.

5. The radio broadcast symbol on the XJN Chart is preferred. It can be distinguished easily from the other radio symbols.

6. On the back of the chart, the radio beacon symbol appearing on the XJN Chart is preferred; the radio broadcast symbol appearing on the XDA Chart is preferred.

7. Range stations are considered the most important radio aids to navigation.

8. In presenting terrain features, pilots prefer the shadient tints of the XDA Chart but with the spot elevations of the XJN Chart.

9. Pilots prefer the presentation of large cities in yellow as shown on the XJN Chart.

10. The names of cities are more easily read on the XJN Chart. This seems to be due to the contrast between the black print and the yellow background of the land area.

11. Pilots prefer the differentiation of roads and railroads as shown on the XJN Chart.

12. Radio information is preferred on both sides of the chart but the symbols should be consistent on both sides.

13. The inbound magnetic headings on the range legs of the XDA Chart and the indication of the "N" quadrants on the XJN Chart were both highly favored.

14. Coastal outlines, cities, roads and railroads, and terminal ranges are desirable on the chart; non-terminal ranges are preferred in a less prominent form.

15. Airways, fan markers, state names and boundaries are of minor importance to the pilot and need not be shown on the chart.

16. The size and scale of the two experimental charts are satisfactory but a new chart combining the best features of both is highly desirable.

### Summary

The major objective of this study was to devise and apply experimental techniques through which data could be obtained and form the basis on which principles of chart construction could be evaluated. Some of these principles seem obvious but until experimental data were available, they remained in the realm of conjecture.

In evaluating the charts, data were obtained from readability tests, an analysis of test items and pilot preferences on a questionnaire. The results indicate that the two experimental charts designed for navigation in high-speed, high-altitude aircraft are superior to traditional charts in presenting information for cross-country missions. On an over-all basis, the experimental charts are not statistically different from one another. However, there are several features on each chart which appear to be highly effective in presenting navigational information to the pilot.

It seems apparent, therefore, that the ideal chart for navigation in high-speed, high-altitude aircraft should include the desirable features of each chart with further experimentation to determine the effectiveness of their interaction.

*Received July 7, 1952.*

## The Relationship between Scotopic Visual Acuity and Acuity at Photopic and Mesopic Brightness Levels \*

J. E. Uhlaner, D. A. Gordon, I. A. Woods, and J. Zeidner

*Personnel Research Section, Personnel Research and Procedure Branch,  
Adjutant General's Office, Dept. of the Army, Washington, D. C.*

The present problem is concerned with visual acuity at various brightness levels, but from a somewhat different point of view from that taken in the usual psychophysical experiment. Classical studies (4, 5) have described visual acuity as a function of brightness level. In such studies, interest is in mean or typical performance. Individual differences are regarded as variability limiting the generality of the findings. In the present problem, we are interested in assessing the possibility of constructing a practical test of night visual acuity. In this regard, we are not concerned with mean performance, but attention is strongly centered on individual differences.

The Personnel Research Section of the Adjutant General's Office has been undertaking, for some time, the development of a practical, predictive test of night visual performance. In early studies carried out at Fort Sill (3) in 1944, and at Camp Blanding (10), the Army Night Vision Tester (ANVT-R2X) was constructed and validated. The instrument is satisfactory from the point of view of reliability and validity, but it has shown itself too cumbersome for general field service.

The present approach of the Personnel Research Section to night vision testing attempts to substitute an acuity test given at mesopic (moonlight) brightness (6.75 log micromicrocrolamberts) for the scotopic (starlight) test. This substitution is desirable because tests of mesopic acuity involve less adaptation time (and hence more rapid testing), less dependence on light-tight testing conditions, and fewer testing personnel. In the practical military situation, these factors might well be critical in determining whether or not a test of night vision could be adapted for extensive use.

A mesopic acuity test may be substituted for a scotopic test, if the relationship between the two tests is shown to be high. Studies reported in the experimental literature indicate that visual acuity scores are correlated at certain brightness levels. The closer the brightness levels tested, the higher has been the correlation reported.

The relationship of acuity at photopic (daylight) and scotopic brightnesses has been investigated in two studies. Uhlaner and Woods, 1951 (7), employing 200 subjects, reported correlations ranging from .19 to .39 between various photopic acuity tests given at 10.02 log  $\mu\mu\text{L}$ . and scores on the Army Night Vision Tester given in the brightness range of 3.51 to 5.26 log  $\mu\mu\text{L}$ . Warden, 1944 (8), however, found biserial correlations of only .02 between scores on the Navy Radium Plaque Adaptometer at 3.94 log  $\mu\mu\text{L}$ ., and scores on a Snellen test given at standard photopic brightnesses. The restriction of range on the photopic variable may partially explain the low correlation attained in this study. The 100 subjects tested all had photopic acuities of 20/20 or better.

Two other studies have been concerned with the relationship of acuity scores taken at adjacent brightness levels in the photopic-mesopic range. L. S. Rowland (6) compared acuities at 10 log  $\mu\mu\text{L}$ ., 7.6 log  $\mu\mu\text{L}$ . and 6.5 log  $\mu\mu\text{L}$ . brightness levels, employing 56 subjects. The tetrachoric correlations between acuities at these levels (computed by the present authors) are: 10 vs. 7.6 log  $\mu\mu\text{L}$ .,  $r = .61$ , 10 vs. 6.5 log  $\mu\mu\text{L}$ .,  $r = .73$ , 7.6 vs. 6.5 log  $\mu\mu\text{L}$ .,  $r = .61$ . Feinberg and Wirt (2) found the intercorrelations of scores of far visual acuity, measured 100 subjects on the Bausch and Lomb Ortho-Rater checkerboard target, at brightnesses ranging from "normal" to 1/33 of "normal" to range from .71 to .90. Gener-

\* The opinions presented in the paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

ally, the closer the levels compared, the higher the correlation attained.

The present problem extends these analyses to a comparison between scotopic visual acuity and acuity at photopic and mesopic brightness levels. From the viewpoint of assessing the practicality of developing mesopic tests to measure scotopic acuity, the present study is crucial. The feasibility of this approach would be demonstrated if indications of sufficiently high correlations could be shown between scotopic and mesopic acuity, and if these correlations were substantially higher than those between scotopic and photopic acuity.

### Method

**Apparatus.** The scotopic measurements were made on the Army Night Vision Tester (ANVT-R2X, 7). This instrument presents a black, two-degree Landolt Ring against a four-degree white background. The intensity of illumination is varied through eight steps of decreasing brightness, by placing filters over the self-luminous radium plaque background. Brightness varies in

these steps between 5.26 and 3.51 log  $\mu\text{P.L.}$  The subject is required to indicate which one of eight positions the break in the ring is facing. Eight presentations of the stimulus are given in random order at each brightness level.

The photopic and mesopic acuity measurements were made on wall charts and on the Bausch and Lomb Ortho-Rater instrument. All tests were conducted in the Pentagon Vision laboratory. This laboratory was standardized in conformity with specifications prescribed by the Armed Forces-National Research Council Vision Committee. The layout at this laboratory is shown in Figures 1 and 2.

The wall charts employed included the Modified Landolt Ring, Army Snellen, Line Resolution, and Quadrant Variable Contrast targets (Figure 3). Except for the Army Snellen, these charts were developed by the Personnel Research Section and were utilized in an earlier factor analysis study of photopic visual acuity.

Photopic and mesopic acuity measurements were also made by means of the Ortho-Rater instrument. The optical system of this instrument presents the test target at an apparent distance of eight meters (1). In the present study, only the far visual acuity adjustment was employed. Control of the voltage input of the Ortho-Rater

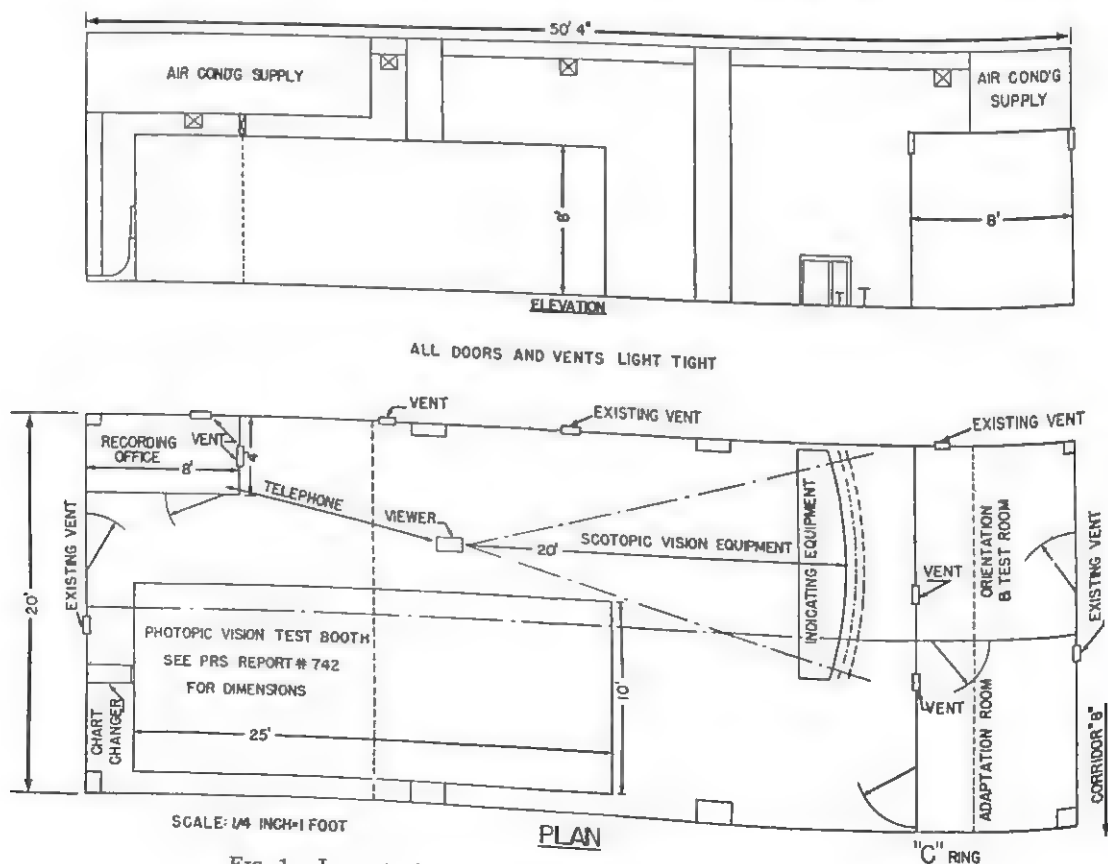


FIG. 1. Layout of the Pentagon Vision Laboratory, floor plan.



subjects were not identical for all the tests. Subjects were selected to sample a wide range of scotopic acuity.

**Procedure.** Testing on (a) wall charts, (b) monocular Ortho-Rater plates, (c) first binocular plates, and (d) second binocular plates occurred in separate sessions. A month intervened between (a) and (b), a week between (b) and (c) and a month separated (c) and (d). Scores on the Army Night Vision Tester had been obtained about a year prior to the commencement of the present study. All testing was conducted with corrected vision on those subjects who customarily wore glasses. The procedure involved in testing with wall and the Ortho-Rater plates will be described separately.

**Wall Charts.** Each subject was dark adapted for 10 minutes in the testing room which was darkened to approximately .001 foot-lamberts brightness. This length of dark adaptation is sufficient to allow valid visual acuity testing to be carried out at the lowest brightness level utilized (Level 8). The tests were observed binocularly in the following order: Modified Landolt Ring, Army Snellen, Line Resolution, and Quadrant Variable Contrast. Testing continued on each subject until he had made three consecutive errors. After the scores were recorded, the light level was adjusted to the next higher brightness level (Level 7); the subject was given an adaptation period ranging from 15 to 30 seconds, and testing again took place in the same order as in Level 8. This procedure was followed for the remaining six levels. The eight levels of illumination employed are shown in foot-lambert and log  $\mu\text{L}$ . See Table 1.

Total time for each subject in each session was approximately 15 minutes.

**Ortho-Rater Tests.** The procedure employed in administering the Ortho-Rater tests was similar to that employed with the wall charts, except that only a single type of test target (letters) was administered at each brightness level.

In the monocular testing, the subject was first

tested with the right-eye target at the lowest level of illumination and then with the left-eye target at the same level of illumination. As in the wall-chart procedure, the level of illumination was raised to the next higher level and the subject was given 15 to 30 seconds to adapt to the higher level. The subject was tested at this level with the right-eye target and then with the left-eye target. This procedure was repeated in the same manner for the remaining six levels of illumination.

Testing at light levels 1, 2, and 8 was omitted from the first binocular test. A preliminary analysis of the monocular data indicated that the targets available did not adequately discriminate between subjects at these levels.

All illumination levels were included in the second binocular test because a new target was employed. With the inclusion of this new target test, no inferences from the monocular data could be drawn as was the case for the first binocular test.

## Results

**Wall Chart.** The relationship between scores on the scotopic test and on each of the wall chart tests at the brightness levels tested is given in Table 2. The Quadrant Variable Contrast test was discarded as it failed to differentiate between subjects. The items of this test were too difficult for the best subjects. Scores on the Army Night Vision Tester were number of correct responses in the 64 presentations constituting the test. Scores on the wall charts were number correct to three consecutive errors. Rank order correlations are shown in Table 2.

The smoothed scores were obtained by fitting through each individual's scores at the various brightness levels, a curve similar in

Table 1  
Levels of Illumination Employed

Level of Illumination	Wall Chart Tests		Ortho-Rater Tests	
	Ft.-Lamberts	Log $\mu\text{L}$ .	Ft.-Lamberts	Log $\mu\text{L}$ .
1	13.5	10.16		
2	3.0	9.51	13.0	10.18
3	.850	8.96	3.0	9.51
4	.080	7.94	.850	8.96
5	.020	7.33	.070	7.85
6	.008	6.94	.020	7.29
7	.003	6.51	.009	6.96
8	.001	6.03	.003	6.51
			.001	6.03

Table 2

Correlations of the Army Night Vision Tester with Raw and Smoothed Wall Chart Test Scores  
N = 15

Level of Illumination	Mod. Landolt		Army Snellen		Line Resolution	
	Raw	Smoothed	Raw	Smoothed	Raw	Smoothed
1	.58	.69	.42	.37	.35	.44
2	.61	.68	.44	.41	.12	.47
3	.59	.65	.47	.41	.53	.62
4	.69	.62	.56	.58	.48	.60
5	.60	.81	.57	.57	.65	.69
6	.62	.87	.61	.51	.53	.62
7	.81	.82	.35	.63	.26*	.58*
8	.57*	.62*	.63*	.63*	.04*	-.01*

\* The relationships implied by these correlations must be accepted with reservation as the mesopic tests upon which they are based showed inadequate differentiation and variance at these levels. Correlations of .51 are significant at the 1 per cent level.

shape to the function which seemed to represent the relationship between acuity and brightness based upon the observations for 15 subjects. These smoothed scores represent an attempt to get scores in which the error variance is minimized. Similar logic is implied in all methods of curve fitting.

*Ortho-Rater.* The relationship between scores on the scotopic test and the Ortho-Rater scores is given in Table 3 below. Scores on the Ortho-Rater are based on the number of rights to three consecutive errors. Best Eye "A" is defined as scores on the eye which gave best acuity on the majority of brightness levels tested. Best Eye "B" is defined as scores on the eye which gave best acuity at

each level. For the first binocular target, testing was carried out only for Levels 3 through 7.

#### Discussion

*Relationship of Scotopic and Higher Brightness Scores.* A trend is found for higher correlations to occur between the wall charts and the Army Night Vision Tester at the lower brightness levels (Table 2). Highest correlation (raw) with the Army Night Vision Tester occurred at Level 7, 6.51 log  $\mu\mu\text{L.}$ , for the Modified Landolt Ring, at Level 6, 6.94 log  $\mu\mu\text{L.}$  for the Army Snellen, and at Level 5, 7.33 log  $\mu\mu\text{L.}$  for the Line Resolution test. The correlations of the Ortho-Rater with the

Table 3

Correlations of the Army Night Vision Tester with Ortho-Rater Test Scores  
N = 16

Level of Illumination	Best Eye "A"	Best Eye "B"	First Binocular	Second Binocular
1	.18	.31	—	.12
2	.59	.59	—	.12
3	.63	.65	.40	.21
4	.68	.75	.54	.38
5	.49	.57	.40	.22
6	.22	.21	.51	.43
7	*	*	.43	.33
8	*	*	—	.35

\* Inadequate differentiation of subjects was shown by the tests at these levels. Correlations of .50 are significant at 5 per cent level.

Army Night Vision Tester show an increase as brightness levels are increased to Level 4 (7.85 log  $\mu\text{L}$ ) and a decrease at higher brightness levels. Highest correlations with the ANVT-R2X are obtained at this level for Best Eye "A" and "B" and for first binocular scores. The second binocular scores show highest correlation at Level 6.

It might reasonably have been expected that scores on the Army Night Vision Tester would correlate most highly with tests administered at the lower brightness levels, i.e., Levels 7 and 8. Failure to obtain this result here may perhaps be explained by the unsuitability of the wall charts and Ortho-Rater targets employed for testing at the lower brightness levels. Correlations appear to increase up to the point where these targets cannot be seen by the subjects.

The alley charts correlate more highly with the Army Night Vision Tester than do the Ortho-Rater plates (Tables 2 and 3). This result may be attributed to the superior acuity distributions at the low brightness levels obtained on the charts. It should be noted that one of the wall charts used the Landolt broken ring design which is similar to the target used in the Army Night Vision Tester. The specificity of the Landolt target may have increased the correlations. Further study should be made to determine whether or not the ring gives high correlations with scotopic tests of other designs.

These results would raise doubt concerning the allegation that scotopic visual acuity scores are too unstable to permit their long-term prediction. In the present study, the Army Night Vision Tester was administered to the subjects a full year before the photopic tests. Despite this time difference, correlations of .60 and higher are found between the Army Night Vision Tester and the mesopic tests.

### Summary

The aim of this study was to determine the correlations among scores on a scotopic visual acuity test and scores on wall charts and Ortho-Rater plates administered at various photopic and mesopic brightness levels. Nineteen subjects were employed, selected to show

a wide range of scotopic acuity scores. The correlations obtained are considered only as indications of relationships due to the small number of subjects employed. Scotopic acuity was measured with the Army Night Vision Tester (ANVT-R2X). Brightnesses ranged from 3.51 to 5.26 log  $\mu\text{L}$ . Mesopic and photopic acuities were measured with various wall chart tests and targets used in a modified Ortho-Rater instrument. Brightness levels ranged from 6.03 to 10.60 log  $\mu\text{L}$ . The main findings of this study are as follows:

1. Scotopic acuity scores showed moderate positive correlations with the photopic acuity scores, and higher correlations with mesopic acuity scores, both for the wall chart tests and the Ortho-Rater plates.

2. The Landolt Ring acuity target shows higher correlations with the Army Night Vision Tester than do the other targets. It is not possible to state whether this result is due to similarity of design of the Landolt Ring and the Night Vision Tester, or to some intrinsic factor of the design itself. In future developmental work on a test of night vision ability, this target should be included as one of the mesopic targets.

3. High correlations with mesopic acuity were obtained in the present study, even though the scotopic test was administered to the subjects a full year before administration of the photopic and mesopic tests. This finding should raise doubts concerning the claim that scotopic visual acuity scores are too unstable to permit their long-term prediction.

4. As a consequence of 1 and 3 above, the practicability of developing a mesopic test of night vision ability is indicated. Such a test would have the following advantages over a scotopic test: shorter adaptation time (hence more rapid testing), less expensive and cumbersome equipment, less dependence on light-tight testing conditions, and fewer testing personnel.

Received June 30, 1952.

### References

1. Armed Forces—NRC Vision Committee, *Manual of Instructions: Armed Services Vision Tester*. University of Michigan, April 1951.

2. Feinberg, R., and Wirt, S. E. Visual acuity in relation to illumination in the Ortho-Rater. *J. appl. Psychol.*, 1947, 31, 406-412.
3. Field Artillery School, *Report on study of night vision*. Fort Sill, Oklahoma, February 1944.
4. Hecht, S. Relation between visual acuity and illumination. *J. gen. Physiol.*, 1928, 11, 255.
5. Lythgoe, R. J. *The measurement of visual acuity*. Medical Research Council, Special Report No. 173, London, His Majesty's Stationery Office, 1932.
6. Rowland, L. S. *Night visual efficiency in illuminations above the level of the cone threshold*. U. S. AAF School of Aviation Medicine, Randolph Field, May 31, 1944 (3551).
7. Uhlaner, J. E., and Woods, I. A. *Studies in night visibility*. Highway Research Board, Bulletin No. 43, 1951.
8. Warden, C. J. *An investigation of motion acuity under scotopic conditions at various retinal positions*. U. S. NRC-CAM, April 1944, Report No. 326.
9. Personnel Research Section, A.G.O. Report 742. *Studies in visual acuity*, 1948.
10. Personnel Research Section, A.G.O. Report 816. *Validation of Army Night Vision Tester*. 30 January 1950.

# The Influence of Increased Positive $g$ on Reaching Movements<sup>1</sup>

A. A. Canfield

Wayne University

A. L. Comrey

University of California at Los Angeles

and

R. C. Wilson

University of Southern California

The pilot of modern high-speed aircraft is faced with many stress situations that were unknown to his predecessors, such as the extreme radial accelerative forces developed when an airplane is maneuvered through a change in direction, as in turns and pull-outs from dives. Popular literature is rich with stories of pilots blacking out (a temporary loss of vision due to decreased blood supply to the eye), suffering sudden displacements of the lower intestines, bleeding at the mouth and ears, etc. Other than the occurrence of blackout and unconsciousness, the latter concomitants of these forces apparently occur rarely, if ever, in practice (1).

Physiologists and medical research specialists, together with engineers, have developed protective clothing called  $g$ -suits in an effort to counteract these radial forces. While these efforts have been successful in elevating the tolerance threshold somewhat, techniques have not been developed to compensate for the tremendous increased effective weight of the body under these increased accelerative conditions. A person exposed to a 5  $g$  accelerative force, by definition, has an effective weight equal to five times his normal weight. Woods *et al.* (2) at the Mayo Clinic centrifuge have shown that it is impossible for a man to rise from his seat under conditions of 5  $g$ . In addition to the problem of general body movement, the increased weight also introduces problems in moving the extremities, as the arms and legs weigh equivalently more. This introduces serious problems for the pilot when he attempts to reach

for and/or manipulate controls under conditions of radial acceleration.

While the effect of these radial accelerative forces on effective body weight is the same irrespective of the direction from which the force is imposed, markedly different physiological effects are associated with them. When the force is applied along the vertical axis of the body from head to seat, blood tends to pool in the abdominal cistern and the lower extremities. This is the commonly experienced positive  $g$  and is the type studied in this paper. When the direction of force application is reversed, blood pools in the head, and this is called negative  $g$ . When the force is applied at right angles to the vertical axis of the body it is called transverse  $g$ , and has generally less serious effects. The tolerance to transverse  $g$  is very high (partially accounting for experiments on the prone position for high-speed aircraft pilots), next highest for positive  $g$ , and low for negative  $g$ .

This research was conducted for the purpose of evaluating the effects of positive  $g$  forces on the speed and accuracy of ballistic reaching movements of the arm. All of the research data were collected on 48 volunteer, but paid, Ss on the human centrifuge located on the University of Southern California campus. Each S had passed a rigorous physical examination before being allowed to participate in the study.

## Experimental Procedure

The 48 Ss were randomly divided into four groups of 12 each. Each group was subjected to three different  $g$  conditions: 1  $g$ , 3  $g$ , and 5  $g$ . All made a ballistic reaching movement with their hand to a target approximately 5" square at a distance of 19" from the starting point. This was a switch on the end of a metal tube which projected toward them at shoulder height and in the midline of the body. From this point they reached at an angle of 35° to the target in each

<sup>1</sup>The research reported in this paper was conducted at the University of Southern California under the auspices of ONR contract N6-ori-77 Task Order III and constituted the doctoral dissertation of the senior author. Dr. Neil D. Warren supervised the research and his kind help and counsel are gratefully acknowledged.

of four positions—up, down, left, and right. The target face was at right angles to the path of movement for all four target positions. The whole target area was well within the maximum working area of the arm as described by Barnes (3).

The switch at the starting point closed a circuit on a standard timer when  $S$  removed his hand. Another micro-switch was placed behind a rubber diaphragm which served as the backing of the target. When  $S$  hit the target surface with his finger, this switch automatically opened the circuit and stopped the timer. Another clock in the circuit started when the starting buzzer sounded and stopped when he hit the target. It was thus possible to derive the following three time scores: the time taken to start the movement, called reaction time in this study, the time taken to make the movement, and the total time which elapsed from the sounding of the buzzer to the completion of the movement.

The face of the target was covered with a sheet of polar coordinate graph paper scribed in intervals of 1 tenth inch. The  $S$ 's preferred finger was covered with a metal cot that terminated in a pin-like point. As the target was struck this point punctured the polar coordinate target sheet. These points indicated the exact location of the strikes. The strikes on the target were considered from two standpoints—the quadrant of the target in which they fell, regardless of the size of target center disparity; and the distance from the center of the target, direction disregarded.

A number of different types of scores were available for comparing  $S$ 's performance at the different  $g$ -levels and target positions such as reaction time, movement time, total time, direction of error, magnitude of error, and the relation between the times and the accuracy of the movements.

Before starting the test trials, each  $S$  was given two indoctrination rides on the centrifuge including a ride at 5  $g$ . If  $S$  desired to continue, his experimental trials were begun. On the first experimental day, each  $S$  spent about fifteen minutes making movements to the target in the position he would encounter on that day. Each  $S$  was also trained to detect the difference between the warning and the reaction buzzers (differing in pitch) and was shown how his responses would be evaluated in the experiment. All  $S$ s were instructed to make the movement as quickly and accurately as possible, and to strike as near the center of the target as they could. During this first day's practice, care was taken to assure that  $S$  make a ballistic movement (4), and not a moving fixation.

Each of the four sub-groups of  $S$ s, 12  $S$ s in each, had different arrangements of target position. Within the framework of the total group, all positions preceded and followed each other an equal number of times. While the target

order was the same for all subjects in any one group, the order of imposed force for members of the group was systematically varied. As a result each  $g$  level and target position preceded and followed each other an equal number of times. These precautions were taken to avoid any experimental error that might result from the serial effects of either  $g$  or target position.

Each  $S$  had two experimental days following the first day of practice. The target was placed in two of the four positions of each of these days, and  $S$  made four consecutive reaching movements for the target at each of the three positive  $g$  conditions used in the experimental—1, 3, 5 (1  $g$  is normal gravitational force, and does not involve centrifuge rotation).

The data of the experiment were 192 movements (4 each for 48  $S$ s) made to each target position for each of the three different  $g$  levels. Only the target position and the radial force imposed were known to vary systematically.

## Results

*Direction of Error.* The observed distribution of the responses in four quadrants of the target demonstrated striking changes as the  $g$  level increased, but the nature of the change varied with target position. Figure 1 shows the number of responses which fell in each of the quadrants for each of the four target positions at the three  $g$  levels.

An examination of Figure 1 shows that at 1  $g$  the movements made upward, to the left and to the right tended to fall in the upper half of the target, and the responses downward fell about equally in the upper and lower halves. The figure also shows that the responses tended to fall on the right side of the target when it was in the "up" and "down" positions, on the left side when in the "left" position, and on the right side when in the "right" position. As the  $g$  level increases, however, the responses moved to the lower half of the target when in the "up," "left," and "right" positions and the upper half when the target was in the "down" position. Similarly, the responses shifted to the right half of the target when it was in the "up," and "left" positions and to the left half when in the "down" and "right" positions.

Table 1 shows the results of Chi Square tests of the distribution of responses between the  $g$  levels for each target position.

Of the 24 values, 16 are significant beyond the 1% level of significance, and in all but

UP	1 g	3 g	5 g
	48   73   121	20   41   61	25   23   48
	33   38   71	36   95   131	49   95   144
	81 111 192	56 136 192	74 118 192
DOWN	1 g	3 g	5 g
	26   63   89	42   38   80	60   47   107
	26   77   103	54   58   112	48   37   85
	52 140	96 96	108 84
LEFT	1 g	3 g	5 g
	82   45   127	45   29   74	24   35   59
	31   34   65	56   62   118	52   81   133
	113 79	101 91	76 116
RIGHT	1 g	3 g	5 g
	38   68   106	58   45   103	37   24   61
	43   43   86	52   37   89	74   57   131
	81 111	110 82	111 81

FIG. 1. Frequency of responses in the various target quadrants by target position and g level.

six instances they are significant beyond the 5% level. The responses clearly tend to move to the nearer and lower quadrants of the target as the g level increases.

*Response Accuracy.* In all of the four target positions the accuracy of the movement was severely impaired by the higher accelerative forces. Table 2 shows the circular errors for the various g levels and target positions.

In all cases the magnitude of the error of movement was larger (significant beyond the 1% level of confidence) at 5 g and 3 g than it was at 1 g. The increase between 3 g and 5 g, however, was not significant for either the "down" position or "right" positions.

The accuracy of movements to the left was significantly poorer at all g levels than those

made downward and to the right.<sup>2</sup> It was also significantly poorer than upward movements except at 5 g where no significant difference was found between the two. Movements to the right, upward, and downward were not significantly different in their accuracy at the 1 g level, but both movements to the right and down were significantly more accurate than reaching upward at the increased g levels.

In general, movements to the left were the least accurate at all g levels, with movements into the other three planes showing no significant difference under normal conditions. Movements to the right and down, however,

<sup>2</sup> Forty-seven of the 48 Ss preferred the right hand for making this type of movement.

Table 1

Chi Square Values from a Comparison of the Obtained Left-Right and Up-Down Splits in Target Strikes at the Various *g* Levels with Each Other\*  
No. of responses = 192

Levels	Target Position							
	Up		Down		Left		Right	
	u-d	l-r	u-d	l-r	u-d	l-r	u-d	l-r
1-3	80.45	13.35	1.70	51.06	65.33	3.09	0.18	17.97
1-5	148.45	1.26	6.79	82.70	107.54	29.44	42.64	19.22
3-5	4.06	8.17	15.62	3.00	4.95	13.06	36.95	0.02

\* Chi<sup>2</sup> of 3.84 significant at the 5% level of confidence; Chi<sup>2</sup> of 6.64 significant at the 1% level of confidence.

are a great deal more accurate than upward movements at the increased *g* levels. Reaching to the right is somewhat more accurate than reaching down at 3 *g*, but no significant difference was found at 1 *g* or 5 *g*.

**Movement Time.** The time required to complete the ballistic movement increased markedly as the *g* level increased for movements upward and to the left, but was less seriously impaired for movements downward and to the right. Table 3 shows the movement times with the target in the four positions and for each of the *g* levels.

The differences in the movement time are significantly higher (beyond the 1% level) for each succeeding *g* condition when the target is in the "up" position. The time required for the movement is similarly, though not as seriously, impaired for movements to the left.

Movements to the downward direction did not show any significant increase in time, and movements to the right were not consistently impaired.

Both the movements made downward and to the right were faster at the increased *g* levels than those made upward and to the left. These differences are all significant beyond the 1% level of significance except the down-left comparison at 3 *g* which is only significant at the 5% level. The only significant difference between the speed of movement at 1 *g* in the various directions was that movements to the right were faster (at the 5% level) than those made upward. No significant differences were found between the speed of reaching to the right or downward.

**Reaction Time.** Previous research (5) has indicated that simple reaction time to both sound and light stimuli increases significantly

Table 2

Means, Standard Deviations, and Standard Error of the Means of the Circular Error Scores\*

No. of subjects = 48

Target Position	<i>g</i> Level					
	1 <i>g</i>		3 <i>g</i>		5 <i>g</i>	
	M	S.D.	M	S.D.	M	S.D.
Up	4.74	1.73	8.38	3.63	9.67	3.84
Down	4.61	2.23	5.84	2.88	6.91	3.72
Left	5.73	2.61	6.97	2.84	9.66	4.57
Right	4.33	1.90	7.15	2.97	7.79	3.82

\* All values presented in this table are given in tenths of inches. Each of the 48 scores from which these values were computed represented the average error score of four responses made at the *g* level and target position indicated.

Table 3

Means, Standard Deviations, and Standard Error of the Means of the Movement Times\*

No. of subjects = 48

Target Position	<i>g</i> Level					
	1 <i>g</i>		3 <i>g</i>		5 <i>g</i>	
	M	S.D.	M	S.D.	M	S.D.
Up	1.35	.295	1.50	.484	2.20	.620
Down	1.31	.337	1.27	.378	1.31	.406
Left	1.33	.306	1.38	.438	1.59	.505
Right	1.28	.323	1.23	.359	1.35	.397

\* All values are given in seconds. Each of the 48 scores from which these values were computed represented the total time taken for four response movements at the *g* level and target position indicated.

Table 4

Phi Coefficients Between Circular Error and Movement Time for the Target Positions by  $g$  Level  
No. of subjects = 48

$g$ Level	Target Position			
	Up	Down	Left	Right
1	-.169	-.416**	-.248	-.416**
3	-.374**	-.390**	-.248	-.500**
5	-.208	-.374**	-.332*	-.627**

\* Significant at the 5% level of confidence.

\*\* Significant at the 1% level of confidence.

at increased  $g$  levels. The term reaction time as used in this study of reaching movements should not be interpreted as a measure of the maximum speed of reaction, but should be considered as a preparation period prior to instigating a movement. This period was significantly longer (beyond the 1% level) for all target positions at 5  $g$  than it was at 1  $g$ , but did not differ significantly between any of the target positions, or correlate significantly with either the accuracy or speed of the movement which followed.

*Relation between Movement Speed and Accuracy.* It was consistently found that longer movement times were associated with greater accuracy. Table 4 shows the Phi coefficients derived from the intercorrelation of each  $S$ 's movement time and accuracy scores at each  $g$  level and target position. The significance of the Phi coefficients were determined through converted Chi Square values.

Inasmuch as speed of movement and amount of error were found to be negatively related, it must follow that since both speed and accuracy were impaired under increased  $g$  conditions, the accuracy would be even further impaired if the same movement time were achieved, and vice versa.

#### Interpretation

*Direction of Error.* The fact that the reaching movements tended to terminate in different quadrants of the target as the  $g$  level increased is attributed to two different sources. The first of these might be termed "experimental error."

Under increased  $g$  conditions, the first response of the  $S$ s to the target would quite frequently fall far below the center of the target. If the trial which followed was a 1  $g$  trial, the first movement was frequently quite high. Both of these types of errors on the first movement (too low at the increased  $g$  conditions and too high at the normal  $g$  condition) were often accompanied by exclamations of surprise. The first movement at 1  $g$  was frequently made in response to the pattern of kinesthetic cues that had been used for moving the arm under the previous atypical weight conditions. This introduced a source of response error that is difficult to judge, but if recognized as a systematic source of error, admits that the error was due to the conditions of the experiment and does not detract from the meaningfulness of the results insofar as the effect of increased  $g$  on movement is concerned.

Second, the accumulation of strikes on the lower and nearer sections of the target suggests that two different types of movement errors occurred. First, the observance of responses on the near side of the target suggests that the initially applied force was insufficient to carry the arm to the intended termination point, and second, the tendency for the strikes to accumulate on the lower half of the target is attributed to an error in judging the required trajectory of movement. Following the terminology of Brown *et al.* (6), the first of these has been called the "negative inertia error," and the second has been termed the "error of downward tendency." As a consequence, one would expect the strikes to fall in the lower half of the target in the "up" position as both errors are acting in the same direction. In addition, since the types of errors are additive in this position, one would anticipate movements to this position to be the least accurate of all. The results verify this deduction. Similarly, the response to the target in the left and right positions would be expected to accumulate on the lower and near section. This is what occurred. With the target in the down position, the errors tend to offset each other, the negative inertia error tending to make them fall on the upper half of the target, and the error of downward

tendency tending to make them fall on the lower half of the target. The results have shown that the responses fell on the upper half at 5  $g$  indicating that the negative inertia error was the more predominant of the two.

**Response Accuracy.** The increase in the error of the movement is attributed to the inadequacy of the normal kinesthetic cues under the increased  $g$  conditions. Possibilities of reduced visual acuity at the 3 and 5  $g$  conditions are highly unlikely in view of previous research findings on perceptual speed ability at these same  $g$  levels (7) and the fact that no  $S$  ever reported gray-out, the preliminary symptoms of blackout.

**Movement Time.** The increase in movement time is attributed to the failure of the  $S$ s to throw the arm with sufficient force to compensate for its increased effective weight. Despite the fact that movements, made at the increased  $g$  levels were in the main shorter (responses falling on the near side of the target) they took longer. The difference in time is hardly within that which might include a shift from a ballistic to a moving fixation movement, and the error pattern reflects no such alteration in method of arm movement.

**Reaction Time.** The increase in reaction time, as defined in this study, is attributed to an increased cogitation period before starting the reaching movement. After the first movement at increased  $g$ ,  $S$ s were immediately aware of the fact that this was a different situation, calling for a different arm thrust. The increase in time between the warning buzzer and the start of the movement is considered an increase in the readiness period taken by the  $S$ s to get better "set" for the ensuing movement.

### Summary

From the results of this research, certain conclusions about the effect of increased positive radial acceleration on reaching movements may be advanced.

1. Both the speed and accuracy of reaching movements at increased  $g$  levels are seriously impaired, the degree of impairment

being roughly equivalent to the amount of force imposed.

2. The kinesthetic cues governing the thrust of the arm under normal circumstances are inadequate to maintain similar accuracy or speed under radial accelerative conditions.

3. Due to the increased weight of the arm and the inadequacy of the normal kinesthetic cues, two types of errors are found, one being the negative inertia error and the other the error of downward tendency.

4. The most favorable location of controls for the pilot of high-speed aircraft, both from the standpoint of speed and accuracy, is to the side of the pilot's preferred hand and below its normal resting point.

5. Emergency controls that might have to be manipulated under conditions of increased positive radial acceleration should be no smaller than two inches in diameter if a pushing motion is required.

Received June 23, 1952.

### References

1. Wood, E. H., Lambert, E. H., and Code, C. F. Do permanent effects result from repeated blackouts caused by positive acceleration? *J. Aviat. Med.*, 1947, **18**, 471-481.
2. Code, C. P., Wood, E. H., and Lambert, E. H. The limiting effect of centripetal acceleration on man's ability to move. *J. Aero. Sci.*, 1947, **14**, 117-123.
3. Barnes, R. M. *Work methods manual*. New York: Wiley and Sons, Inc., 1944.
4. Hartson, L. D. Analysis of skilled movements. *Personnel J.*, 1932, **11**, 28-43.
5. Canfield, A. A., Comrey, A. L., and Wilson, R. C. A study of reaction time to light and sound as related to increased positive radial acceleration. *J. Aviat. Med.*, 1949, **20**, No. 5.
6. Brown, C. W., Ghiselli, E. E., Jarrett, R. F., and Mimium, E. W. *Speed and accuracy of spatial location in the prone position*. Aero. Med. Laboratory, Eng. Div., Air Materiel Command, Wright Field, Dayton, O. Serial No. MCREXD-694-4H, E. O., 1948, 694-17.
7. Canfield, A. A., Comrey, A. L., Wilson, R. C., and Zimmerman, W. S. *The effect of increased positive radial acceleration upon human abilities (Part II: Perceptual speed ability)*. U. of S. Calif., Dept. of Psychol., Office of Naval Res. Contract N6 ori 77, Task Order III, Report No. 4, 1948.

## Applied Psychology in Action

*Editor's Note:* An announcement of this new feature of *J. appl. Psychol.*, including a plea for psychologists in business and industry to send in suitable material was sent to about 40 psychologists on the firing line early in February, 1953. As of the beginning of April, not a single response had been received.

As noted in the April issue, this new feature will be continued only long enough to determine whether or not our readers desire such a section and whether or not psychologists will take the time and trouble to submit suitable copy.

### Job Supervision of Young Workers

The following is extracted from a Report of the Technical Committee on Supervision of Young Workers, Bureau of Labor Standards, U. S. Department of Labor, February, 1953, composed of: Chairman, Mrs. Margaret F. Ackroyd, Chief, Division of Women and Children, Department of Labor, Providence, Rhode Island; Fanny G. Buss, Standard Oil Company, Cleveland, Ohio; Mrs. Mary Cooper, Hutzler Brothers Restaurant, Baltimore, Maryland; Jane F. Culbert, Vocational Advisory Service, New York City; Gilbert David, The Prudential Insurance Company, Newark, New Jersey; James Forster, DeKalb Agricultural Association, Inc., DeKalb, Illinois; Harry Gladstine, The Washington Post, Washington, D. C.; Dr. Dale Harris, Institute of Child Welfare, University of Minnesota; Mrs. Bernice Heffner, American Federation of Government Employees, Washington, D. C.; Kathryn-Lee Keep, Department of Labor and Industry, Erie, Pennsylvania; R. Bruce Neill, James Monroe High School, Fredericksburg, Virginia; Clyde L. Schwyhart, Caterpillar Tractor Company, Peoria, Illinois; Thomas E. Walsh, Amalgamated Clothing Workers of America, Troy, New York; Benjamin C. Willis, Superintendent of Schools, Buffalo, New York; and Mrs. Gertrude Folks Zimand, National Child Labor Committee, New York City.

*"What Should the Supervisor Know About Youth?"* The Committee believed that the core of the task of getting better supervision of young workers is to help supervisors of youth to be more interested in and better understand the basic characteristics of youth—their capabilities, their problems, their at-

titudes, and their needs. The responsibilities and the attitudes of work supervisors necessarily center largely about a concept of workers as adults. Nine-tenths of the Nation's workforce is indeed past 20. Yet almost every industry and business has at least a small component of young beginners. That youth are not yet adult is a truism, but too often not fully understood by the man or woman whose responsibility it is to help youth give good work performance and grow up to be good workers.

In attempting to define the characteristics of youth of significance to a work supervisor, attention was focused upon youth of about 14 to 18 years of age. The Committee believed, however, that a description of this midadolescent group would be useful in understanding older youth on the job as well. The Committee was exceedingly grateful to its member, Dr. Dale Harris, for preparing in advance of the meeting an analysis of the characteristics of youth in their midadolescence with special reference to those characteristics likely to be significant in job situations. Bringing the practical job experience of various Committee members to bear on Dr. Harris' contribution, the Committee developed the following description:

*Youth—A Period of Adjustment.* The supervisor must first of all realize that adolescent boys and girls are in transition from childhood to adulthood, and that this stage in a person's development may be a difficult period of personal adjustment. This transitional stage has no precise age limits, but is defined by psychologists as beginning at roughly 12 to 14 years and continuing to 21

or 22 years. The midadolescent period of about 14 to 18 years of age is normally the period of greatest stress.

The major areas of adjustment that are the primary concerns of teen-agers are usually considered to center about four factors: (1) how to be attractive to the opposite sex, (2) problems of family relations which result from their attempts to emancipate themselves from parental control, (3) for those still in school, anxiety over school achievements, (4) concern with vocational plans—though this may often be vague and unrealistic.

Of great significance to those who supervise the early work experience of adolescents is the youth's concern to be considered somebody, a person of importance to himself and others, with a place in the world and a contribution to make.

Of equal significance to supervisors is youth's own insecurity about the emerging responsibilities and challenges of adulthood and how to act to realize them. They do not like to admit these insecurities, but they are nevertheless there. Young people therefore reach out for security which in large part they attempt to find by tying closely to a group of their own age. They seek the approval of that group, and conformance to its standards becomes very important to them.

In the adolescent's desire to be grown-up, he sometimes has difficulty in accepting the authority of adults. This adolescent 'revolt' expresses itself in various ways—including the display of immoderate behavior, language or dress.

*Basic Characteristics.* There is of course, a tremendous range of differences among individuals at adolescence as at any age. No two are indeed alike. However, the basic characteristics of the adolescent stage of development which the supervisor of young workers needs to be aware of, can in general be described as follows:

*Physical Maturity.* Most girls will be physically matured by the age of 14; a good many boys are still immature at this age. Consequently girls are much more likely to appear mature and socially poised than boys their own age.

Strength is closely related to physical maturity. Boys gain 50 per cent in actual muscle volume during adolescence, girls much less. Sexually immature individuals at this age will lack considerably in strength and endurance. There are great differences in strength between physically mature and immature boys of the same ages.

Adolescents can mobilize much energy on demand, but not all youth are able to maintain sustained output. Furthermore, growth in muscle volume may lag behind growth in stature; a youth may not be as strong as he first appears.

Many adolescents of this age have yet to learn how to achieve a balance between physical needs for rest on the one hand, and social interest and needs on the other.

Physical health is good; the period is characterized by little illness.

Basic motor skills, such as speed of movement, reaction time, and coordination are fully developed although not necessarily fully trained.

*Intellectual Maturity.* Intellectual stature has just about been reached; measurable increments of intelligence after fifteen are much less significant than those which occur from ages ten to fifteen. Many older adults fail to recognize that the average adolescent of sixteen and seventeen has achieved sharpness of intellect and a heightened readiness to learn.

Even though he may be 'bright,' the adolescent is limited in 'judgment.' Though he lacks experience, he resents being talked down to and being considered unable to solve problems.

The adolescent's ability to think abstractly is well developed, which leads him to seek reasons based on principle. This sometimes makes him appear argumentative.

Adolescents are able to evaluate their own behavior, and actually they engage in a great deal of self-criticism. They are often quite sensitive to blame, though they may not seemingly admit failure when criticized. They may be easily discouraged.

Many adolescents exhibit a great deal of intellectual and emotional 'questing'—a vague

longing for something unknown. This makes them receptive to emotional appeals to loyalty, integrity, self-sacrifice.

*Interests and Attitudes.* The day-to-day behavior of the midadolescent will show considerable vacillation between the developing interests of maturity and the interests of the younger child, though the adolescent himself frequently will reject the less mature interests after returning to them temporarily. Frequently adults see this as unpredictability and unreliability.

Many adolescents are characterized by considerable idealism and a sense of altruism, and at the same time by snobbishness and a feeling of superiority.

The critical capacities of the adolescent may extend to others, so that he appears to be highly intolerant. Combined with idealism, this characteristic leads him to seek perfection in adults, and he may feel let down when they fail to measure up to his expectations. These attitudes may carry over into his relationships with his work supervisor.

Adolescents frequently have a strong desire to do well, and to get ahead. Although

they are somewhat vague about specific goals in life they want 'to know where they are going.'

*Social Behavior.* Much of the social behavior of adolescents is characterized by apparent contradictions which upon closer investigation are found to be more apparent than real.

There is a strong desire to be treated as individuals; there is also a strong desire to conform to the standards set by young people their own age.

On the other hand, there may be much deliberate imitation of the attitudes and actions of adult associates, especially of those they admire.

There is a strong need to be independent; there is also a strong need to be dependent.

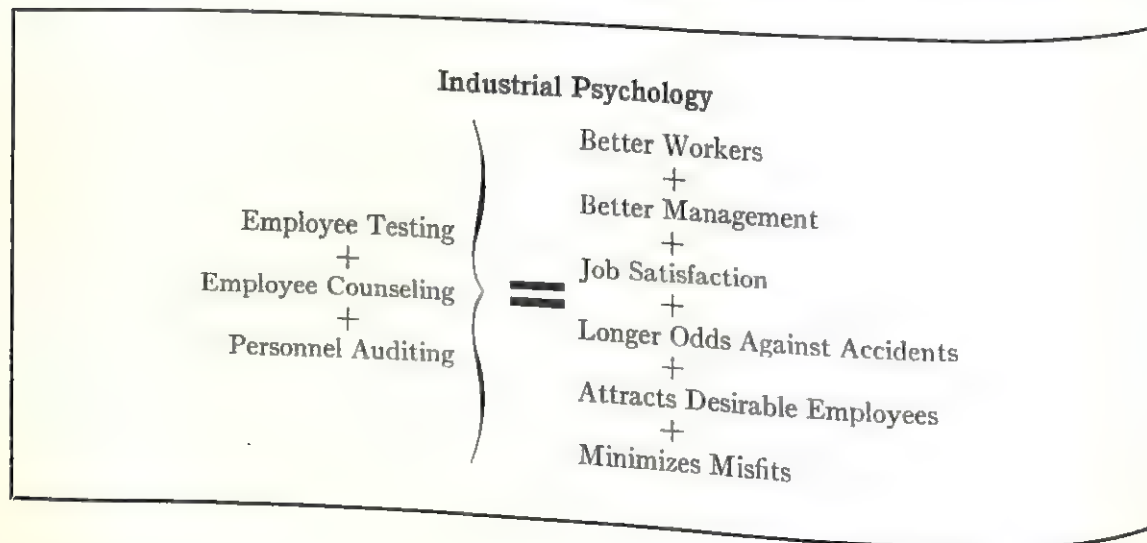
Language is ostentatiously colorful, slangy and emotional. Adolescents' use of profanity or obscenities may actually be an attempt to appear sophisticated and mature.

The adolescent is typically group-minded; he wants to 'belong,' and will respond readily to the idea of teamwork."

### Personnel Psychology in a Steel Company

The work of Personnel Psychologist George M. Hill at the Armco Steel Corporation, Middletown, Ohio, was featured in the February

19, 1953 issue of *The Iron Age*, pp. 61-62. The following interesting chart was included in the article:



## Book Reviews

Kephart, Newell C. *The employment interview in industry*. New York, McGraw-Hill, 1952. Pp. 277. \$4.50.

The book covers the main items regarding content and method of the employment interview and a lot of other material about employment procedures. It is on this point that the author might be criticized, namely including so much other material in a book entitled *Employment Interview*. It would seem from the discussion that the interviewer gives tests and visual examinations, diagnoses mental maladjustments and interprets all available predictors of the criterion. To be sure this is all pertinent to the process of hiring and the discussion of these aspects is sound. Actually there is scarcely enough material on the face-to-face aspect of the employment interview to make a respectable book and the author presumably did what any of us would have done under the circumstances.

The initial chapter sets the place of the interview with reference to usual employment procedures. Then follow chapters about the content of the interview and some "how to" aspects. The first of these involves knowledge of the job with due reference to the *Dictionary of Occupational Titles* and various blanks devised by the War Manpower Commission. It is helpful to have some of this WMC material in a handy form. The next chapter deals with evaluating past experience by means of job families, vocations, and Volume Four of the *Dictionary*. Tests of intelligence and of motor ability are considered. The author cautions against inferring intelligence from the interview alone but does indicate some things such as vocabulary or sentence structure manifested in the interview that might give some indirect evidence.

A chapter on personality includes specific items of behavior that might be observed during an interview. It also discusses clinical symptoms and syndromes. An interesting suggestion is attempting to find a job for an applicant with serious personality deviations who may, nevertheless, adjust to some kind

of work. This is a commendable acceptance of industry's social responsibility. In connection with physical demands of the job there is due emphasis on vocational possibilities for persons with physical disabilities. Emotional maturity is mentioned as especially important for leadership jobs and some specific interview questions are suggested which might bring out emotional maturity.

The last two chapters deal more specifically with the mechanics of the interview—what the reader would anticipate from the title of the book. One considers preliminary preparation—the actual environment of the interview and the application forms. There is a tabulation of items involved in a considerable number of application forms which might help someone in devising a form of his own. With reference to the actual conduct of the interview there is emphasis on the avoidance of bias and of stereotyped methods. The patterned interview is recommended on the basis of some experimental studies of reliability and validity. The over-all conclusion appears to be that the interview is needed to supplement tests because the latter do not get at everything needed in the job and do not have perfect validity. The reviewer is disposed to add that as we perfect objective personality tests the importance of the interview may ultimately decrease. A final topic is the importance of giving the applicant adequate information about such things as hazards, working conditions and possibilities in the job. This aspect might very profitably receive more stress in the discussion.

The book is moderately well documented with references at the end of each chapter to a few pertinent experimental studies. The general level is fairly elementary except for an occasional mention of something like multiple correlation and the book evidently is designed for the person without much technical background. There are a lot of wise cautions for an interviewer, such as not being misled by a good talker. There are also a lot of hints as to what kind of questions to ask in order to bring out indirectly some aspect of personality. Finally, there is a wholesome

emphasis on the social aspects of the employment program and of industry's responsibility for the over-all adjustment of the worker.

Harold E. Burt

Ohio State University

Wechsler, David. *The range of human capacities*. Second Edition. Baltimore: Williams and Wilkins Co., 1952. Pp. 190. \$4.00.

This is a revision of Wechsler's 1935 book of the same title. New chapters on productive operations and span of life have been added and the chapter on the effect of age has been enlarged. There has been minor rewriting in many parts of the book and it has been completely reset.

The purpose of the book is still "to show that human variability, when compared to that of other phenomena in nature is extremely limited, and that the differences which separate human beings from one another . . . are far smaller than is ordinarily supposed."

In pursuit of this objective, Wechsler gathers data concerning human capacities (defined to include such diverse measures as temperature, height, reaction time and intelligence) and compares the score of the lowest person with the highest person within the *normal* population. Normal population is arbitrarily defined as excluding one-tenth of one per cent of the total population at each extreme. The comparison effected is in terms of the *range ratio* which is simply the highest score or value divided by the lowest. Wechsler notes that these range ratios are small (i.e., less than 5) and asserts that they are in the nature of natural constants. A hierarchy of range ratios is postulated extending from about 1.30:1 in the case of linear traits (such as stature and arm length) to about 2.5:1 in the case of perceptual and intellectual abilities. It is implied that the "real" upper limit is probably the growth constant  $e$  (2.7182) and an attempt is made to show that the orderly hierarchy of ratios is a function of the number of factors involved in the various human capacities.

In the new chapter on productive operations, data are reviewed on employee pro-

ductivity and it is concluded that a ratio of 2.00:1 expresses the difference between the best and poorest workers. This is interpreted as indicating that one cannot expect much from selection techniques and need not be concerned about uniform pay vs. sliding pay scales.

The chapters on Length of Life, Exceptions, and the Burden of Age, while interesting, are extraneous to the major theme. For example, expected life span gives very large range ratios at whatever period of life the ratios are computed and Wechsler concludes that life span is either not a capacity or that the data are badly contaminated.

In the chapter on Genius and Deficiency Wechsler embraces the *theory of critical differences* to explain both ends of the continuum. After a given quantitative change, he asserts, qualitative distinctions appear. The ability to see new relationships might be such a change at the upper end of the scale. This qualitative change, it is maintained, accounts for differences which greatly exceed the "mere 50 IQ points" which separate the genius and the idiot from the average. In his last chapter on the Meaning of Differences, however, Wechsler apparently forsakes this line of thinking and returns to the refrain that if the range ratios yield small numbers, then the differences which separate men are inconsequential.

An appendix on mental measurement and one containing his basic data complete the book.

To this reviewer, the book suffers from three major confusions. *First*, it seems obvious from the frequent reference to social significance, democracy, rule by the elite, etc. that Wechsler feels that to believe in democracy one must demonstrate that all people are really equal in everything from body temperature to test scores. This, of course, is a confusion of *value* judgments with descriptive physical and psychological statements. It is perhaps a common confusion but should be deplored all the more for that fact. *Second*, while Wechsler knows the rules for measurement and the necessary prerequisites for making meaningful ratios, he apparently does not apply these rules to all of his data

and reports range ratios for Mental Ages, IQ's and number of items correct in a vocabulary test. It is obvious that these ratios have no significance since they do not have known, meaningful zero points and equal units. *Third*, while Wechsler recognizes that words like "small" and "large" are judgmental words whose meaning is not clear without a set of references, he persists in saying that because range ratios can be expressed in "small numbers" (arbitrarily defined), they are "small." And, since they are "small," the differences between people are "small" and *this* has great social significance. In other words, "small" is defined in one context and applied in another where it carries added meanings.

The work is further marred by misinterpretation of the data and a running series of errors. In the chapter on productive operations, for example, ten range ratios are introduced as evidence: 1.73, 2.00, 2.04, 2.10, 2.30, 2.53, 2.55, 2.57, 2.83, 3.00. From this array it is concluded that the range of productivity is ". . . not more than 2.5:1 and generally more nearly 2.0:1" (!). No evidence is ever given for the repeated statement that *e* is the probable upper limit of the ratio. Reference is made to figures and tables which disagree with the text (e.g., Figure 5, Table 7); a significant claim is made about modes in the data, one of which is nonexistent, etc. The invitation to check the results by recalculating the data in the appendix is not reassuring. In a cursory examination of these data the reviewer found fifteen cases of considerable error. Many of these errors also appear elsewhere in the book. Either the data have been misprinted, the original range ratios miscalculated, or both. In two cases the data are patently impossible.

Wechsler's basic notion of the range ratio offers interesting and intriguing research ideas when confined to the kinds of data to which it legitimately applies. In the present context its potential value appears to be buried in a host of confusions and irrelevancies. There appears to be no more need for the second edition of this book than there was for the first edition.

James J. Jenkins

University of Minnesota

Division of Occupational Analysis, United States Employment Service. *Dictionary of occupational titles, second edition*. Washington: United States Government Printing Office, 1949. Volume I, Definitions of Titles, Pp. xxviii + 1518, \$4.00. Volume II, Occupational Classification and Industry Index, Pp. xxvi + 743, \$2.50.

Users of the DOT should welcome the Second Edition because it provides them with more occupational information in more usable form than did the early edition. Volume I contains the job definitions including those from the old Part I, the various supplements, and additional new definitions. The appendices from the original edition (Glossary; Index of Commodities to assist in classifying Sales Personnel; Occupational Titles Arranged by Industry) have been moved to Volume II of the Second Edition. Other readily apparent changes are the introduction of a double alphabetic scheme of presenting definitions and a considerable simplification of the reference techniques. The main alphabetic listing presents every job and occupational title by straight letter alphabetization. This is a desirable change over the former word alphabetic listing which was more troublesome to users because of the numerous compound and multi-word titles. For example, in the original edition CELLAR WORKER preceded CELLARMAN, while in the new edition the order is reversed. Within the main listing are indented sub-listings of job definitions most closely related to the base definition; thus, users are saved the time of locating these definitions throughout the volume and can more readily compare the different definitions. The variety of reference phrases found in the first edition are now reduced to the words "see" and "see under." Teachers and others providing instruction in DOT usage will join the user in approving these changes.

Changes not immediately apparent have also been effected. Coverage has been expanded within the professional occupations as well as within several industrial categories. Codes now accompany all job definitions previously referred to classification titles. Four of the so-called grouping title defini-

tions have been eliminated and type-of-work classifications introduced for many laboring jobs. Such changes, coupled with the double alphabetic listings, mean that thousands of additional jobs are now readily codable, which is a marked contrast to the cumbersome multiple reference processes required with the first edition.

Glimpses of other possible improvements are found by comparing the Second Edition with the former publications for such classifications as CHEMICAL ENGINEER, ELECTRICAL ENGINEER, and CHECKER (clerical) III. Redundancy, overlapping, and repetitious classifications have been substantially eliminated with no significant loss of occupational information. It is regrettable indeed that publication was not delayed until a host of similarly needed changes were made—also until the remaining grouping title definitions, so frustrating to users, were eliminated.

A serious complaint against the original edition was the amount of training time required to achieve proficiency in its use. Here, again, the Second Edition is an improvement—experience having already shown that training time is about one third less.

The general format of the Occupational Classification in Volume II remains essentially the same with the different titles being readily identifiable so that users may locate those with definitions in Volume I directly. Users will be pleased to note the elimination of the LABOR, PROCESS classifications.

This reviewer strongly feels that the kinds of improvements found in the Second Edition should have been extended throughout many additional occupational areas.

Alan M. Kershner

Personnel Research Center, Inc.,  
Arlington, Va.

Prasad, Kali. *Fatigue and efficiency in textile industry*. Lucknow, India: Univ. Lucknow Press, 1950. Pp. iii + 34. Rs. 1/8 or 2s. 3d.

This is one report in a continuing series of research studies begun in 1947 in the Swadeshi Cotton Mills, India. Four operations

were studied in detail. Output data were analyzed by hours, days, months, and shifts. Psychophysical tests were administered to some employees, and a questionnaire was administered to a sample.

It is difficult to evaluate this book properly because of cultural differences in the degree of psychological sophistication between India and the United States. Further, this book is not the final report of the entire series of studies.

From the viewpoint of United States psychology, this study is weak in several important respects. Fatigue is defined as "a condition . . . caused by activity in which the output produced by that activity tends to be rather poor, and the degree of fatigue tends to vary directly with the poverty of output." It appears to this reviewer that this definition also covers "monotony," for example. It might be better to concentrate on variations in output, and forget the fatigue.

The mill had 9,404 employees, but most data refer to extremely small samples, i.e., 4, 8, 33, etc. The "critical level" or "ideal performance" of each worker was based on a one-half hour sample of his output. These samples are not only too small, but also subject to disturbing variables such as the "Hawthorne effect." It is not always clear whether data were "experimentally" collected, or taken from routine records. The significance of much of the data is not clear.

Since Indian psychology and economy are both rather new, it is possible that this is an important study in India. The present book is not particularly valuable to Americans. Perhaps the final report of the whole study will be useful.

Harold F. Rothe

American Hospital Supply Corporation,  
Chicago, Illinois

Lauer, A. R. *Learning to drive safely*. Minneapolis: Burgess Publishing Co., 1949. Pp. 145. \$2.25.

This manual conveniently and precisely presents a well-conceived driver training course for the course administrator, the driving instructor, and the student. In the words

of the author: "This manual is the result of twenty years' study of drivers' aptitudes, habits, abilities, and disabilities, in addition to ten years' experience in teaching drivers and instructors of driving at Iowa State College. Every step outlined has been carefully tested and evaluated for difficulty, order of presentation, and usefulness. . . ."

Duties and responsibilities of all connected with the course from administrator to student are specifically prescribed, including such details as solicitation and payment of fees and *principles and procedures to protect students* and equipment. The course material itself consists of ten basic units of instruction, which may be covered in ten or more lessons. Each unit contains an introduction directed to the student, an outline of skills to be mastered, a few reference readings, a list of questions, and a student's report form. A valuable appendix contains specific suggestions for handling classes, suggested administrative forms, psychophysical and psychological tests, a list of equipment needed, and a list of films and visual training aids.

While the manual is directed to a non-professional audience, there are two items of particular interest to psychologists. First, the author considers the development of the proper attitudes toward driving as a most important part of the course. He stresses the need for the course to begin with reading and classroom discussion and exercises in order "to broaden their interest in good driving and the philosophy of safety education." Secondly, the appendix does contain a short section on the interpretation and use of psychophysical and psychological tests. One hopes that the users of the tests contained in the manual will seek advice from persons qualified to interpret tests results. The reviewer finds a statement encouraging this action conspicuously lacking.

In summation, the manual should be a tremendous aid to the school administrator planning to establish a driver training course or seeking to improve an existing course. Its primary value to psychologists, as well as to all citizens, will not come from reading the manual, but will accrue from the lowering of

the accident rate as more and better formal training courses are established.

Stanley E. Jacobs

Department of the Army,  
Washington, D. C.

Shostrom, Everett, and Brammer, Lawrence.  
*The dynamics of the counseling process.*  
New York: McGraw-Hill Co., 1952. Pp.  
xvi + 213. \$3.50.

A book designed to meet the present needs of counseling should certainly deal with, *among other things, problems of definition*, real or apparently conflicting theories and practices, foundations in general psychology, especially learning and motivation, and the role of counseling in its various settings. This book was apparently so designed. Unfortunately, it falls short of the mark at almost every point.

Its core is a description of the "self-adjustive" approach which represents one more attempt to synthesize the Minnesota and Chicago positions. It is Rogerian in its major features but tries to make room for testing and informational procedures. It is defined as ". . . counseling which assists the client to become more self-directive and self-responsible" (p. 2). While it may be more appealing to state a definition in terms of goals rather than in terms of operations, it is probably less scientifically useful. Fortunately, this does not affect the discussion of actual methods which is fairly well done and describes procedures that have been followed for some time in the better counseling centers.

Concerning the synthesis of opposing points of view, the authors furnish their own best criticism: "It is this middle-of-the-road stand, taken by so many counselors, which has created more confusion than clarification in counseling methodology" (p. 4). For despite their diagrams and denials, they are, if not on a continuum (which probably does not exist), certainly in between.

In the process, they tilt at the usual windmill, directive counseling, and bandy about the usual tired, emotionally-toned, invidious comparisons in which the good (i.e., Rogerian, permissive, or self-adjustive methodology) is

characterized as "democratic" as in the following: "It would appear that the basic assumptions of democracy and those of client-centered therapy are one and the same . . ." (p. 10). Their *bête noire* is Williamson's fifteen-year-old text on counseling.

The authors' attention to a systematic basis on learning theory for their method is commendable, but the reviewer was puzzled by their use of John Dewey's 1933 book as the principal (and almost exclusive) core of the presentation. The names of Hull, Guthrie, and Tolman do not appear, and regrettably little is made of the cited works of Dollard and Miller, Mowrer, and Shoben.

The Stanford Guidance Study is presented as an example of research on counseling. In it, "Feeling tone . . . was the criterion used to evaluate the effectiveness of counseling" (p. 41). Their conclusions are based on differences between ratings of such feelings; differences are *described* as significant, but no statistics are presented to document this.

Methodologically, they have accepted several doubtfully valid notions. For example, they say, "It is assumed that clients are *capable* of selecting their own tests" (p. 74). Also, they suggest that, "Perhaps if counselors would concentrate less on the limitations of students and more on the limitations of test data, the quality of guidance would improve" (p. 29). Perhaps! But this is a rather naive criticism of one of the best developed aspects of counseling. And it is difficult to see how the authors could have failed to see the implications of their statement that, "The *only* (italics added) indicators that anxiety has been reduced are the client's feelings expressed toward himself and the counseling services" (p. 151).

While the foreword makes much of the fact that this book conceives of counseling as an integral part of education, the actual discussion of the role of counseling in colleges and universities is limited to six pages of quite superficial description of needs. The training of counselors is dealt with in fourteen lines which emphasize the value of electrical recordings; these presumably give the trainee a knowledge of rather than a knowledge about counseling.

This book left the reviewer with one overriding impression: that it is *undigested*. The authors are obviously enthusiastic and ambitious; they are aware of the major needs of counseling; they have written well and made their points forcefully. Yet the total product is, regrettably, unsatisfactory.

John W. Gustad

University Counseling Center,  
University of Maryland

Guetzkow, H. (Ed.). *Groups, leadership and men*. Pittsburgh: Carnegie University Press, 1951. Pp. ix, 293. \$5.00.

This book presents progress reports of five years (1945-1950) of contract research in Human Relations sponsored by the Office of Naval Research. These twenty reports are revisions of papers given at a mutual stock-taking conference which was held at Dearborn, Michigan in September, 1950. Psychologists predominate among the contributors which also include sociologists, political scientists, economists, and journalists, all of whom were members of the research teams involved in the undertaking.

The book is divided into the three main sections suggested by the title. About half of the space and total number of reports are included in the first section, which deals with research on the behavior of groups. R. B. Cattell introduces this section with the formulation of methodology and basic concepts. Following this discussion are a number of reports by leading members of several University of Michigan research centers. Among the subjects treated are components of group morale, the effects of communication on non-conformists, workers' loyalties to union and management, and factors making for group productivity. The section is concluded with Margaret Mead's paper on research in contemporary cultures.

The second section deals with problems of leadership. Topics discussed include: the influence of the group in determining leadership style, the relation of the follower's personality to the leader, and leadership effectiveness at the production level.

In the final section, which is concerned with individual behavior, the psychological reader

is brought back to terra firma. New light is cast on traditional problems of measuring motivation, the relationship of verbal behavior to the reasoning process, and the advantages of neuropsychiatric screening.

Some technical detail has been omitted from the original reports in order to make them suitable for a wider readership. However, references at the end of most chapters have been supplied to aid the more curious social scientists in following up these overviews.

A service is rendered the reader by John G. Darley who has contributed introductory and concluding chapters designed to give perspective and integration to an assortment of competent but somewhat discontinuous reports of on-going research. The reader who is more concerned with practical military application is accommodated by a discussion at the end of the book, and those interested in securing contract subsidy for their projects will find the appendix helpful.

In general, the content of the book is more of a prologue to a new social psychology than a report of substantial achievement. The atmosphere is one of more problems raised than solved, and the predominant theme is "further research needed." But there is a healthy respect for the canons of scientific method by the seasoned researchers who have contributed to this volume. Although the reports deal with path-finding in new territories, the projects generally involve problems that are reduced to testable hypotheses. The generalizations are for the most part tentative and limited to the data actually involved, with a notable absence of intuitive and sweeping conclusions. Nevertheless, the reviewer concurs with Darley in the expressed need for more synthesis and higher order generalization, since the visions gained by this exploratory work may tend to be obscured by the trees. Another source of uneasiness also made explicit by Darley is the insufficient consideration of the role of abilities and interests as determinants of group behavior. Since psychology has had considerable success in these areas, even a prologue may benefit from the past.

This volume is a useful source of supple-

mentary reading for students in social psychology, and is of particular interest to the multitude of social scientists now in the employ of various Armed Forces Human Resources programs. It has a wider appeal than most technical publications of in-service military groups, for the emphasis is on basic research which the Navy has so far-sightedly underwritten. Also, it is a good example of what may emerge from the large-scale institutional research which has become another sign of our times.

Abraham S. Levine

Bureau of Naval Personnel,  
Washington, D. C.

Curran, C. A. *Counseling in Catholic life and education*. New York: The Macmillan Co., 1952. Pp. 462. \$4.50.

This book is a new approach to counseling in a number of ways. It is new in combining an accurate knowledge of modern counseling techniques as they have developed in the fields of psychology and education in America with the Thomistic and Aristotelian concepts of the virtues. In addition, it definitely relates religion and counseling together.

In its technical presentation, this book clearly distinguishes counseling from guidance and so opens the way for the use of both types of relationships with persons who come for help. Curran defines counseling as "a definite relationship where, through the counselor's sensitive understanding and skillful responses, a person objectively surveys the past and present factors which enter into his personal confusions and conflicts and, at the same time, reorganizes his emotional reactions so that he not only chooses better ways to reach his reasonable goals, but has sufficient confidence, courage, and moderation to act on these choices." Elsewhere he has defined guidance as "a relationship in which a person equipped in a particular field supplies pertinent facts to an immediate personal need. Guidance readiness occurs," he says, "when a unique convergence of events in a person's personal life makes a particular kind of information far more meaningful at a given point in life than it would be at any other period." In this book, the discussion of coun-

seling presupposes adequate instruction and information obtained from teaching or guidance and treats these questions only to explain more clearly some aspect of counseling or reasons for a particular kind of counselor method.

The book is made up of five parts. The first part includes some important recent developments in counseling. The second section delineates the process of personal integration as it occurs in counseling from the point of view of the person who comes for counseling. This is of special interest for beginning counselors who may not have an experiential grasp of a counseling series as the person goes through it. The experienced counselor, too, may find, as did the present reviewer, numerous considerations not treated in other books on counseling. Part III presents the counselor's side by unfolding the skill of the counselor as it varies throughout the different phases of counseling. This section gives a detailed exposition of the counselor's skill in each of the five stages of counseling described: establishing the relationship, initiating counseling dynamics, later phases, and the final stages of counseling. A chapter on skills with children is also included. Most of this part is given over to the different methods of the counselor's responses so that deepest content of a person's statements may be objectively unfolded and reflected. The detailed excerpts from actual interviews are

exceptionally suitable for a careful study of the counselor's skill. Part IV, the approach to counseling, is directed to increasing counselor sensitivity to counseling atmosphere, to disguised expressions of counseling need, and to the *ways in which informational and guidance roles may facilitate counseling*. It includes also a chapter on group discussion and group counseling. The concluding chapter is an integration of counseling with religion. This is especially valuable since both counseling and religion aim at aiding a person to be more at peace with God and himself, happier, and more able to lead an independent, responsible, achieving life.

While this book has a definitely Catholic application as its title indicates, yet the title could be misunderstood and therefore misleading. The content of the book would readily be shared, in this reviewer's opinion, by chaplains of any denomination as well as by psychologists, psychiatrists, and educators and, in fact, by any persons who have active religious beliefs and convictions and wish to see how such a religious point of view can be integrated with modern methods of counseling and guidance. A special merit of the book is that it achieves this integration without losing any of the rigor and exactness of a careful scientific study.

Robert J. Sherry

*Hq. Army Field Forces,  
Fort Monroe, Virginia*

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Understanding that boy of yours.* Melbourne S. Applegate. Washington, D. C.: Public Affairs Press, 1953. Pp. 52.
- Rudolf Pintner, in memoriam.* Seth Arsenian, Ed. Washington, D. C.: Gallaudet College Press, 1953. Pp. 63.
- Innovation, the basis of cultural change.* H. G. Barnett. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 462. \$6.50.
- Getting along with people.* Eugene J. Bengt. New London: Bureau of Business Practice, National Foremen's Institute, Inc., 1952. Pp. 29. \$2.5.
- Practical psychology.* Karl S. Bernhardt. Second edition. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 337. \$3.75.
- Psychoanalytic theories of personality.* Gerald S. Blum. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 219. \$3.75.
- Social factors related to job satisfaction.* Research Monograph No. 70. Robert P. Bullock. Columbus: Bureau of Business Research, Ohio State University, 1952. Pp. 105. \$2.00.
- Human relations I. Cases in concrete social science.* Hugh Cabot and Joseph A. Kahl. Cambridge: Harvard University Press, 1953. Pp. 273. \$4.25.
- Human relations II. Concepts in concrete social science.* Hugh Cabot and Joseph A. Kahl. Cambridge: Harvard University Press, 1953. Pp. 333. \$4.75.
- Phantasy in childhood.* Audrey Davidson and Judith Fay. New York: Philosophical Library, 1953. Pp. 188. \$4.75.
- Marriage, morals and sex in America.* Sidney Ditzion. New York: Bookman Associates, 1953. Pp. 440. \$4.50.
- Statistics in psychology and education.* Henry E. Garrett. New York: Longmans, Green and Co., Inc., 1953. Pp. 460. \$5.00.
- The human senses.* Frank A. Geldard. New York: John Wiley and Sons, Inc., 1953. Pp. 365. \$5.00.
- The intimate life.* J. Norval Geldenhuys. New York: Philosophical Library, 1952. Pp. 96. \$2.75.
- Solitude and privacy.* Paul Halmos. New York: Philosophical Library, 1953. Pp. 181. \$4.75.
- Writing clinical reports.* Kenneth R. Hammond and Jeremiah M. Allen, Jr. New York: Prentice-Hall, Inc., 1953. Pp. 288.
- Selected case problems in industrial management.* Paul E. Holden and Frank K. Shallenberger. New York: Prentice-Hall, Inc., 1953. Pp. 318. \$3.75.
- The changing culture of a factory.* Elliott Jaques. New York: Dryden Press, 1952. Pp. 341. \$4.25.
- The Vienna circle.* Victor Kraft. New York: Philosophical Library, 1953. Pp. 209. \$3.75.
- A history of psychology in autobiography, Vol. IV.* Herbert S. Langfeld and Edward G. Boring, Editors. Worcester: Clark University Press, 1953. Pp. 372. \$7.50.
- Human factors in air transportation.* Ross A. McFarland. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 830. \$13.00.
- Children in play therapy.* Clark E. Moustakas. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 213. \$3.50.
- Freudian psycho-antics, fact and fraud in psychoanalysis.* Maurice Natenberg. Chicago: Regent House, publishers, 1953. Pp. 101. \$2.00.
- How to improve classroom testing.* C. W. Odell. Dubuque: Wm. C. Brown Co., 1953. Pp. 156. \$3.00.
- The Wechsler-Bellevue scales, a guide for counselors.* C. H. Patterson. Springfield, Ill.: Charles C. Thomas, publisher, 1953. Pp. 146. \$3.75.
- A manual for administrative analysis.* John M. Pfiffner and S. Owen Lane. Dubuque: Wm. C. Brown Co., 1953. Pp. 88. \$2.50.
- The best years of your life.* Marie Beynon Ray. Boston: Little, Brown and Co., 1952. Pp. 300. \$3.95.
- Administering the elementary school.* Reavis, Pierce, Stullken and Smith. New York: Prentice-Hall, Inc., 1953. Pp. 384.
- Cases of public personnel administration.* Henry Reining. Dubuque: Wm. C. Brown Co., 1953. Pp. 142. \$3.00.
- The Soviet impact on society.* Dagobert D. Runes. New York: Philosophical Library, 1953. Pp. 202. \$3.75.
- Science and human behavior.* B. F. Skinner. New York: The Macmillan Co., 1953. Pp. 461. \$4.00.
- The stepchild.* William Carlson Smith. Chicago: University of Chicago Press, 1953. Pp. 314. \$6.00.
- New York television.* Dallas W. Smythe. Urbana: National Association of Educational Broadcasters, Gregory Hall, 1952. Pp. 108.
- Readings in learning.* Lawrence M. Stolurow, editor. New York: Prentice-Hall, Inc., 1953. Pp. 512.
- Your child and his problems.* Joseph D. Teicher. Boston: Little, Brown and Co., 1953. Pp. 302. \$3.75.
- Student deferment in selective service.* M. H. Trytten. Minneapolis: University of Minnesota Press, 1953. \$3.00.
- Motivation and morale in industry.* Morris S. Viteles. New York: W. W. Norton and Co., Inc., 1953. \$7.50.
- Research in the international organization field. Some notes on a possible focus.* Richard W. Van Wagenen. Princeton: Center for Research on World Political Institutions, Princeton University, 1952. Pp. 78.

- New means of studying color blindness and normal foveal color vision.* Gordon L. Walls and Ravenna W. Mathews. Berkeley: University of California Press, 1952. Pp. 172. \$2.50.
- Philosophy and psycho-analysis.* John Wisdom. New York: Philosophical Library, 1953. Pp. 282. \$5.75.
- Success in psychotherapy.* Werner Wolff and Joseph A. Precker. New York: Grune and Stratton, 1952. Pp. 196. \$4.75.
- Research into the causes of feeble-mindedness. A symposium.* Utica: State Hospitals Press, 1952. Pp. 37.
- Industrial development at home and abroad—problems and prospects.* Financial Management Series No. 101. New York: American Management Association, 1952. Pp. 28. \$1.25.
- Interracial practices in the YMCA.* National Study Commission on Interracial Practices in the YMCA. New York: Association Press, 1953. Pp. 48. \$1.00.
- Personnel administration and the development of the personnel staff.* Personnel Planning Project No. 20-6-51. Director, Personnel Planning, DCS/P Headquarters, Air Training Command, 1952. Pp. 106. Gratis, limited number of copies available.
- Preparing employees for retirement.* Personnel Series No. 142. New York: American Management Association, 1951. Pp. 27. \$1.25.
- Operating problems of personnel administration.* Personnel Series No. 144. New York: American Management Association, 1952. Pp. 40. \$1.25.
- Practical approaches to supervisory and executive development.* Personnel Series No. 145. New York: American Management Association, 1952. Pp. 42. \$1.25.
- Spotlighting the labor-management scene.* Personnel Series No. 147. New York: American Management Association, 1952. Pp. 43. \$1.25.
- Theses in the social sciences.* UNESCO. New York: Columbia University Press, 1952. Pp. 236. \$1.25.
- Comparative survey on juvenile delinquency.* United Nations. New York: Columbia University Press, 1952. Pp. 132. \$1.00.
- Traffic in women and children.* United Nations. New York: Columbia University Press, 1952. Pp. 43. \$40.

# Journal of Applied Psychology

VOL. 37, No. 4

AUGUST, 1953

## Socio-Psychological Factors in Industrial Morale: II

Raymond E. Bernberg

*Los Angeles State College*

Previously reported research (1) compared the predictive ability of different tests of morale for performance indicators in the work situation. The tests of morale were based upon current concepts; viz.: group morale (GM); employee attitude toward the company (CM) (i.e., acceptance of the formal organization by its members); rating of the supervisors by the workers (S); and self-rating of morale by the workers (SM). Performance indicators used were absences, tardiness, short time absences, medical-aid unit visits, and merit rating. The results indicated a general failing of all tests of morale to have

made of the predictive value the other three tests of morale had for it.

Table 1 contains the Pearson product moment intercorrelations of the four measures based upon 890 cases. The results of the multiple regression analysis are in Table 2. It appears that GM with a correlation of plus .67 with SM has contributed almost the entirety of the multiple R of plus .69.

This result gives much validity to the concept of morale as a group phenomenon as measured by the group morale test. This, of course, is based upon the assumption that the collective opinions of the workers themselves

Table 1  
Intercorrelations of the Tests of Morale

	GM	CM	S	SM
GM	—	.77	.51	.67
CM	.77	—	.48	.47
S	.51	.48	—	.49
SM	.67	.47	.49	—

much predictive value for any of the performance indicators.<sup>1</sup>

An interesting result showed up when SM (a thermometer scale requesting the worker to check along a 0 to 100-degree parameter with verbal referents the degree to which he agreed with the proposition that his work group had high morale exemplified by their working together as a well organized team) was taken as a criterion and a determination

Table 2  
Beta Weights and Multiple R with SM as Criterion

	1 (GM)	2 (CM)	3 (S)
$\beta$	.71	-.16	.16

$$R_{0.122} = .69; \sigma R_{0.122} = .02.$$

are an adequate criterion for appraising morale in the work situation.

The group morale test (2) is a projective type paper and pencil test using the direction of perception technique of attitude measurement. It is based upon content derived from six determiners discussed by Krech and Crutchfield (4). They are: 1) positive goals; 2) satisfaction of accessory needs; 3) sense of advance toward goals; 4) level of aspiration and achievement; 5) time perspective; and 6) feelings of identification, solidarity, and involvement. There are 34 items in the test, all equally weighted.

The question which presented itself next was: If this test with its total content has

<sup>1</sup> This conclusion is in conflict with the multiple correlation of .71 between six objective indices of efficiency and morale scores. See Giese, J. G., and Ruter, H. W. An objective analysis of morale. *J. appl. Psychol.*, 1949, 33, 421-428.—*Editor.*

such a high relationship with the collective opinion of workers concerning their cognition of morale in their work group, which items with their specific content comprise the maximal possible prediction for the test as a whole? This is an analytical question which in essence asks: What do the workers mean by morale? This of course concerns only the relationship of the content of the items of the test and the criterion.

Gengerelli (3) has developed a technique of analysis whereby one may reduce a large battery of tests in their relation to a criterion, to a number of sub-tests which provide the maximal prediction of the criterion. This method describes the whole battery in terms of the smaller sub-set and provides a regression equation to express the total correlation

Table 3

Intercorrelations of the Four "Factors"  
(Items) and SM

Item	Item				
	SM	3	9	11	18
SM	—	.17	.14	.32	.82
3	.17	—	.16	.00	-.14
9	.14	.16	—	.12	-.34
11	.32	.00	.12	—	.17
18	.82	-.14	-.34	.17	—

matrix. He states that this method yields the sub-tests as empirical tests which might be considered as "factors" of a sort. This provides an immediate practical solution to answer the analytical question posed above.

The SM measure was taken as the criterion and the 34 items of the GM test as the battery of tests for the analysis. Because of the large number of cases, 100 of the 890 were selected randomly for determining the intercorrelations between the items (tetrachoric  $r$ ) and the correlations between the items and the criterion (bi-serial  $r$ ).

The results of the analysis produced 4 factors which as a sub-set have a multiple R plus .96 with the criterion. Table 3 shows the intercorrelations between these items and with the criterion. Table 4 indicates their beta weights.

Table 4

Beta Weights and Multiple R with SM as Criterion

	1 (Item 3)	2 (Item 9)	3 (Item 11)	4 (Item 18)
$\beta$	.218	.419	.130	.820

$$R_{0.1234} = .96; \sigma R_{0.1234} = .09.$$

These four "factors" are the following items of the test: Item 3. "Scientific studies show that in groups such as yours, if you wanted to hold a big party, picnic or other type of friendly gathering, you would not care to invite: (a) 60% of your work group; (b) 35% of your work group. Item 9. "Statistics show continually that the increase in group production is a result of: (a) group effort toward step-up; (b) individual effort. Item 11. "Recent industrial studies have shown that the following percentage of workers in a group such as yours gives a good amount of attention to ways of getting ahead: (a) 60%; (b) 90%. Item 18. "A recent poll of workers in groups such as yours found that workers got a lot of satisfaction from working together: (a) infrequently; (b) frequently."

The possible importance of these findings need not lie in the measurement of morale per se. It appears that it would be wise in developing and controlling work groups to consider the following factors: (1) satisfaction of men from working together; (2) increase in production as a result of group effort; (3) intimacy of workers with each other beyond as well as in the work surroundings; and (4) the individual level of aspiration in getting ahead.

Received October 3, 1952.

## References

1. Bernberg, R. E. Socio-psychological factors in industrial morale: I. The prediction of specific indicators. *J. soc. Psychol.*, 1952, 36, 73-81.
2. Bernberg, R. E. The direction of perception technique of attitude measurement. *Int. J. Opin. Attit. Res.*, 1951, 5, 397-406.
3. Krech, D. and Crutchfield, R. S. *Theory and problems of social psychology*. New York: McGraw-Hill, 1948.
4. Gengerelli, J. A. A method of analysis in which the factors are empirical tests. *J. Psychol.*, 1952, 33, 159-174.

## Predicting Success in Dental School

Wilbur L. Layton

*Student Counseling Bureau, University of Minnesota*

During the past several years, a trend in the field of testing has been indicated by the establishment of nation-wide testing programs in which applicants for admission to certain schools, mainly professional schools, are tested in centralized programs. All participating colleges or schools then use the resulting test scores in the process of selecting and admitting students.

However, a test battery which is valid as a selection device at one institution may not be valid at another institution. Rarely do the administrators of the national testing programs publish validity data pertaining to specific institutions. Hence most institutions participating in a national program are doing so blindly unless they evaluate the test battery in their own situation.

Since the fall of 1946, the School of Dentistry at the University of Minnesota has participated in the testing program sponsored by the Council on Dental Education of the American Dental Association by testing freshman classes at the beginning of the school year. During the first five years of this testing program, an attempt was made to determine the relationship between various tests administered in the program and success in Dental School. In 1951 the Council on Dental Education decided that the test battery currently used was adequate and should be used by dental schools to select students for their freshman classes. However, little evidence was available locally to indicate the usefulness of the battery for selection purposes.

Consequently, a study was designed that would correlate pre-dental grades and the test data available through the national program with grades earned by dental school freshmen who entered the University of Minnesota Dental School in the years 1946 through 1949. First year grades in the dental school were the primary criteria for each of the four classes. For the freshman class entering in 1946 four year grades were also available as

criteria. Separate grades in freshman courses in physiological chemistry, anatomy and prosthetics were also used as criteria for all four classes studied. These data were punched into Hollerith cards to facilitate analysis.

Table 1 presents the means and standard deviations for the data which were available for the classes entering in the various years 1946 through 1949. Table 1 indicates that the tests included in the battery varied from year to year. The battery was stabilized in 1949. This battery is the one currently (1952) being used in the national program.

The ACE is the American Council on Education psychological examination. It is one of the most widely used scholastic aptitude tests. It provides three scores,—a linguistic or *L* score, a quantitative or *Q* score, and a total score. Only the total score was used in this investigation.

*The Survey of Object Visualization* is similar in content to the Revised Minnesota Paper Form Board. *The Survey of Natural Sciences* is a 90-item test measuring facts and applications of principles in biology, chemistry and physics. The total raw score was used in the analyses.

*The Carving Dexterity test* is a chalk carving test developed by the Council on Dental Education. The examinee uses a knife, ruler and pencil to carve a piece of chalk which measures approximately  $3\frac{1}{4}$ " in length and  $\frac{5}{8}$ " in diameter to correspond to an illustrated figure. The carvings are graded by judges on the basis of objectives such as: "flatness of surfaces," "clean-cutness of angles," "symmetry" and "accuracy of reproduction." The test supposedly measures finger-knife dexterity as well as spatial visualization.

*GED No. 3* is the USAFI test of General Educational Development test three, "Interpretation of Reading Materials in the Natural Sciences." It measures speed of reading and comprehension of passages in the natural sciences. In 1947, 1948 and 1949 only the

Table 1

Means and Standard Deviations of Criteria and Predictive Variables for Groups Entering the Dental School in 1946, 1947, 1948 and 1949

	1946 N = 81		1947 N = 80		1948 N = 84		1949 N = 88	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.
ACE	131.9	20.9	136.3	15.9	127.8	18.5	134.1	17.5
Carving	13.3	3.1	11.4	2.6	9.9	2.3	11.8	2.9
Survey of Obj. Vis.	29.2	7.4	28.9	6.4	32.1	6.0	30.9	7.5
Survey of Sciences	*		55.4	8.0	55.4	6.9	55.3	9.0
GED No. 3	55.5	12.1	34.0	5.3	34.2	5.2	34.7	5.2
GED No. 1 A	16.6	4.0	16.0	4.0	15.6	3.6	*	
GED No. 1 B	82.2	11.4	83.2	11.1	83.1	9.4	*	
Mich. Vocabulary								
Phys. Science	20.1	3.8	*		*		*	
Biol. Science	18.4	4.9	*		*		*	
Peterson Word Dex.	35.8	8.6	*		*		*	
Pre-dent. HPR**	1.4	.4	1.8	.4	1.8	.4	1.7	.5
Frosh HPR**	1.5	.4	1.7	.4	1.7	.4	1.4	.5
Four year HPR**	1.6	.3						
Anatomy HPR**	1.6	.5	1.5	.5	1.5	.5	1.4	.7
Phys. Chem. HPR**	1.5	.6	1.5	.8	1.8	.7	1.3	.8
Prosthetics HPR**	1.9	.4	1.9	.3	1.7	.4	1.8	.5

\* Not given.

\*\* HPR =  $\frac{\text{honor points}}{\text{number of credits}}$  (A = 3; B = 2; C = 1; D, F = 0 honor points).

first 45 out of the total 90 items were given.

*GED No. 1* is the USAFI test of General Educational Development test one, "Correctness and Effectiveness of Expression." It consists of a section on spelling and a section in

which corrections are to be made in punctuation, words and phrases in a connected text. In Table 1, *A* is the score on the spelling section of the test; *B* is the score on the total test.

*The Michigan Vocabulary* is the Michigan Vocabulary Profile test. The two parts "physical science vocabulary" and "biological science vocabulary" were included in the test battery in 1946.

*The Peterson Word Dexterity* test measures specifically the extent to which the student knows the meaning of certain suffixes and prefixes and measures "dexterity" at manipulating word parts and word meanings.

Coefficients of correlation between the predictive indices and total first year grades and grades in the three first year courses in the Dental School were computed for all four groups. Four year grades for the group entering in 1946 were correlated with the predictive indices also.

A multiple correlation and a regression equation were computed for the 1949 group.

Table 2

Coefficients of Various Predictive Indices with Freshman and Cumulative (4 year) Honor Point Ratios in the Dental School for the 81 Students in the Class Entering in 1946

Predictor	Freshman Grades	Four Year Grades
Freshman HPR	—	.83
ACE Total	.15	.09
GED Reading	.23	.10
GED No. 1 A	.40	.27
GED No. 1 B	.40	.24
Word Dexterity	.29	.25
Mich. Vocab. (Phys. Science)	.29	.12
Mich. Vocab. (Biol. Science)	.37	.19
Survey Obj. Vis.	.14	.49
Carving	.22	.31
Pre-dental HPR	.40	.11

Table 3

Correlation Coefficients of the Various Predictive Indices with Freshman Honor Point Ratio in the Dental School for Classes Entering in 1946 Through 1949

Criterion: Freshman HPR	ACE (Total)	Survey Obj. Vis.	Carv- ing	Survey of Sciences†	GED No. 3*	Pre- dent. HPR	No. of Cases
1946	.15	.14	.22	—	.23	.40	81
1947	.13	.38	.20	.43	.37	.27	80
1948	.11	.29	.21	.17	-.01	.40	84
1949	.34	.29	.18	.43	.34	.42	88

† Not given in 1946.

\* Only  $\frac{1}{2}$  of test given to 1947, 1948 and 1949 groups.

This group was selected because it was given the tests currently being given in the battery. These tests presumably are the most valid on a nationwide basis.

Table 2 presents the coefficients of correlation between the predictive indices and freshman and four year grades for the class entering in 1946. With an N of 81 the standard error of  $r$  if the true  $r$  is zero is .11.

Freshman grades correlated rather highly ( $r = .83$ ) with total four year grades. It is interesting to note that the correlations of the predictive variables with four year grades are smaller than those with freshman grades alone. However, the *Survey of Object Visualization* and *Carving Dexterity* tests are more highly related to four year grades. This may be due to the small number of courses requiring these skills that are given during the first two years of the curriculum. The last two years of the curriculum are heavily loaded with practicum and clinic courses which require these skills of the students.

Table 3 presents the correlation coefficients for various predictive indices for each of four

classes with freshman honor point ratio in the dental school. Tests given in Table 3 are the tests which are currently used in the National Testing Program plus the pre-dental honor point ratio.

In three of the four years (1947 being the exception) the pre-dental honor point ratio was the best predictor of freshman honor point ratios in the dental school. The group entering in 1947 was heavily loaded with veteran students who had taken their pre-dental work before the war. The difference in motivation for these students in their pre-war school work and their post-war school work may account for the relatively lower correlation in the 1947 group. The findings of Hansen and Paterson (1) that there is a "striking increase in post-war scholastic achievement as compared with pre-war scholastic achievement of the same students" (all veterans) tend to support this interpretation. It is interesting to note that the ACE score did not correlate highly with freshman honor point ratio in the first three classes studied, but did for the group entering in 1949. The *Survey of Sciences* test cor-

Table 4

Intercorrelations of the Variables for the 88 Students in the Class Entering the Dental School in 1949

	ACE Total	Survey of Object Vis.	Carving Dexterity	Survey of Sciences	GED No. 3	Pre-dental HPR
Fresh. HPR	.34	.29	.18	.43	.34	.42
ACE Total		.20	.17	.37	.43	.41
Survey of Object Vis.			.43	.29	.27	.02
Carving Dexterity				.15	.17	.11
Survey of Sciences					.46	.28
GED No. 3						.24

Table 5

Correlations of Honor Point Ratios in Oral Anatomy 50-51-52, Prosthetics 50-51-52, Physiology 58-59 with Five Predictive Measures for Four Classes of Dental School Freshmen

	1946† N = 81	1947 N = 86	1948 N = 90	1949 N = 85
a. Oral anatomy 50-51-52 HPR with:				
1. Pre-dental HPR	.37	.26	.26	.41
2. ACE Total	.12	.08	.10	.37
3. Survey of Obj. Vis.	.35	.32	.37	.34
4. Carving	.42	.30	.31	.24
5. Survey of Sciences	—	.18	.17	.38
6. GED No. 3	.08	.11	-.04	.27
b. Prosthetics 50-51-52 HPR with:				
1. Pre-dental HPR	.28	.11	.05	.26
2. ACE Total	.00	.20	.09	.43
3. Survey of Obj. Vis.	.21	.29	.13	.16
4. Carving	.29	.11	.01	.01
5. Survey of Sciences	—	.49	.14	.51
6. GED No. 3	-.15	.09	-.14	.04
c. Physiology 58-59 HPR with:				
1. Pre-dental HPR	.22	.25	.41	.34
2. ACE Total	.05	.04	-.06	.14
3. Survey of Obj. Vis.	.07	.11	.35	.19
4. Carving	.18	.22	.36	.33
5. Survey of Sciences	—	.05	.14	.10
6. GED No. 3	.14	.50	.03	.39

† Survey of Sciences Test was not given to the Freshman Class in 1946.

relates higher with first year honor point ratio for the 1947 class and for the 1949 class than for the 1948 class.

The fluctuation of the coefficients of correlation from year to year may be due to the differences from year to year in the means and standard deviations for the variables and hence differences in the composition of the classes. Changes in the curriculum could also account for the variability of the correlation coefficients. Also, because of the small number of cases studied in the various classes, this variation may be only random variation.

The class entering in 1949 was used for the computation of intercorrelations of the several variables, a coefficient of multiple correlation and a regression equation.

Table 4 presents the intercorrelation of the variables studied.

The regression equation and coefficient of multiple correlation were computed by the Doolittle method. Of the six predictor variables only two, *Survey of Sciences* and pre-dental HPR, yielded significant Beta coefficients.

The coefficient of multiple correlation obtained was .54.

The regression equation is:

$$\hat{y} = .02 X_1 + .39 X_2 - .29.$$

$X_1$  = Survey of Sciences total raw score;  
 $X_2$  = Pre-dental HPR;

$\hat{y}$  = Predicted freshman year honor point ratio.

Table 5 presents the coefficients of correlation between five predictive indices and honor point ratios in oral anatomy, prosthetics and physiological chemistry for the years 1946 through 1949.

The coefficients of correlation presented in Table 5 also fluctuate by variable and by year. This variation may be due to random sampling error, actual changes in the makeup of the classes or changes in the curriculum.

### Discussion

It would appear that the five tests currently retained in the national testing program of the Council of Dental Education of the American Dental Association are not highly related

to grades earned by students in the University of Minnesota Dental School. The ACE test and the carving test are not predictive of first year honor point ratios in the course areas where they might be expected to have some relationship. Weiss (3) has reported results very similar to those presented here. He found that pre-dental grades and part of the *Survey of Sciences* test gave moderately high correlations with theory and technic grades. Peterson (2) has reported some validity coefficients for this test battery. The correlations which he reports are consistently higher than the ones obtained in the present study. Hence, the present study and that of Weiss illustrate the need for local validation of tests used in national testing programs. Tests which appeared fairly good on a nationwide basis did not show up well in the two dental schools studied.

These studies also show the variability in coefficients of correlation one can obtain when several groups are studied. This means that findings based on one group or a nationwide study should be applied with caution in working with another group for counseling or admission purposes. The use of test data should be tempered by careful consideration of all other available data.

Received August 25, 1952.

#### References

1. Hansen, L. M. and Paterson, D. G. Scholastic achievement of veterans. *Sch. & Soc.*, 1949, 69, 195-197.
2. Peterson, S. Forecasting the success of freshman dental students through the aptitude testing program. *J. Amer. Dent. Assn.*, 1948, 37, 259-265.
3. Weiss, I. Predicting academic success in dental school. *J. appl. Psychol.*, 1952, 36, 11-14.

## Prediction and Practice Tests at the College Level<sup>1</sup>

Scarvia B. Anderson

*George Peabody College for Teachers*<sup>2</sup>

Attempts to predict college "success" have been so numerous in recent years that it is a rare freshman indeed who has not been subjected to a testing program of one sort or another. The battery of tests used in the present study may be distinguished from hundreds of similar batteries chiefly in that it includes two very easy practice tests which do not seem *prima facie* to belong with the other more conventional placement tests. The questions which we shall attempt to answer here are "What value do these practice tests have in predicting freshman grade point ratio?" and "How does this value compare with the predictive values of the other tests in the battery?"

The original decision to use practice tests in the freshman testing program at George Peabody College for Teachers was based upon the belief that within a group of entering freshmen there would be wide variability in "test-wiseness" and ensuing diminished reliability of test scores. We reasoned, on the basis of observations in previous years, that some of the freshmen would have had no experience with objective tests and that many of the students would not be familiar with the use of special pencils and separate machine-scored answer sheets.<sup>3</sup>

It was felt that a preliminary testing situation would serve two useful purposes: it would aid the students to become more skillful technically, and thus to concentrate their later efforts on the subject matter of the placement tests rather than on the mechanics; and it would perhaps relieve some of their tensions and anxieties.

<sup>1</sup> Dr. Julian C. Stanley, George Peabody College for Teachers, offered helpful suggestions and criticisms during the preparation of this article.

<sup>2</sup> The author is now associated with the Naval Research Laboratory.

<sup>3</sup> It was determined later from the freshmen who were tested that only one-half had previously used separate answer sheets and special pencils, and that the majority of these had used them only once in a State testing program.

The tests selected for practice tests were *Otis Quick-Scoring Mental Ability Tests: Beta Tests, Forms Cm and Dm*. These tests were designed for grades 4-9, so it was thought that very few, if any, of the Peabody freshmen would have had any experience with the tests for at least four years.

For each of the forms, Cm and Dm, the time limit stipulated in the *Manual of Directions* is thirty minutes. However, since the students' scores were not to be compared with any norms set up by Otis, time limits of 15 minutes (on the Cm) and 10 minutes (on the Dm) were set. An informal set of instructions appropriate to the groups was substituted for the standardized instructions. These instructions were, of course, identical for groups taking the tests in different rooms. The method of marking the separate answer sheets was explained; the students were told to guess if they wished, but emphasis was placed on the statement that they should not waste time guessing<sup>4</sup>; and the groups were informed that the individual results on the Otis tests would be kept confidential and would not go on their records or to their advisers.

The tests making up the regular freshman battery were given after the practice tests and in this order:

1. *American Council on Education Psychological Examination, 1949 College Edition*, Linguistic Tests (ACE-L); and Quantitative Tests (ACE-Q).

2. *Cooperative English Test C2, Form S*, Vocabulary (C2-V); Speed of Reading Comprehension (C2-S); and Level of Reading Comprehension (C2-L).

3. *Cooperative English Test A: Mechanics of Expression, Form T (Eng A)*.

Each individual's total scores and subscores on these three tests, expressed graphically as "normally" spaced percentile ranks based upon local norms, were furnished to the fresh-

<sup>4</sup> Test scores were corrected for "chance."

Table 1

Intercorrelations and Beta Weights Involved in the Prediction of First Quarter  
Grade Point Ratio from Freshman Tests  
(N = 136)

	0 GPR	1 Cm+Dm	2 ACE-L	3 ACE-Q	4 C2-V	5 C2-S	6 C2-L	7 Eng A
1 Cm+Dm	.48		.75	.62	.59	.67	.57	.67
2 ACE-L	.44			.51	.84	.78	.72	.70
3 ACE-Q	.39				.40	.56	.53	.53
4 C2-V	.25					.69	.65	.60
5 C2-S	.42						.90	.57
6 C2-L	.38							.51
7 Eng A	.54							
Beta Weights		.124	.265	.032	-.427	.104	.075	.413

man advisers as an aid in placement and counseling.

### Results

Grade point ratios for one quarter and for three quarters were computed for the students who were included in the study.<sup>5</sup> Excluded from the one-quarter analysis were freshmen who carried less than eight quarter hours and from the three-quarter analysis were freshmen who carried less than 24 quarter hours. A few students, who arrived late or who for other reasons were not present for the first administration of the tests, were given the tests later under less desirable con-

ditions, and it was felt that their test scores were perhaps not comparable with those of the original group. They too were omitted from the study.

The correlations between the seven test scores and between those test scores and GPR for one quarter and for three quarters are shown in Tables 1 and 2.

Note that for the first quarter the correlation between Cm+Dm and GPR exceeds all other correlations with GPR except that of Eng A; for three quarters,  $r$  between Cm+Dm and GPR exceeds the correlations of GPR with ACE-Q, C2-V, C2-S, and C2-L. Among these validity coefficients, however, only the differences between Cm+Dm and C2-V for one quarter and between Cm+Dm and ACE-Q for three quarters are significant beyond the one per cent level of confidence,

<sup>5</sup> GPR was computed, letting A+ = 12 points per quarter hour, A = 11 points, A- = 10 points, B+ = 9 points, and so on down to D- = 1 point, F = 0 points.

Table 2

Intercorrelations and Beta Weights Involved in the Prediction of Three-Quarter  
Grade Point Ratio from Freshman Tests  
(N = 119)

	0 GPR	1 Cm+Dm	2 ACE-L	3 ACE-Q	4 C2-V	5 C2-S	6 C2-L	7 Eng A
1 Cm+Dm	.50		.75	.62	.59	.68	.59	.65
2 ACE-L	.51			.49	.83	.76	.72	.71
3 ACE-Q	.30				.40	.57	.53	.52
4 C2-V	.45					.68	.66	.61
5 C2-S	.43						.90	.57
6 C2-L	.41							.50
7 Eng A	.55							
Beta Weights		.242	.013	-.115	.064	-.082	.166	.367

when the significance test for the difference between  $r$ 's with a common criterial array is applied.<sup>6</sup>

The regression equations shown below do not meet the requirements of the most economical predictive equations.<sup>7</sup> Rather, the beta weights were computed as a means of studying and comparing the contributions of the various tests to the prediction of GPR. The computation of the beta weights was carried out by a modified Doolittle method.<sup>8</sup>

For one quarter  $r_{0.1234567} = .62$ ,<sup>9</sup> and the predictive equation, using standard scores, is

$$z_0' = .124z_1 + .265z_2 + .032z_3 - .427z_4 \\ + .104z_5 + .075z_6 + .413z_7.$$

1 refers to Cm + Dm, 2 to ACE-L, 3 to ACE-Q, 4 to C2-V, 5 to C2-S, 6 to C2-L, and 7 to Eng A.

The largest beta weight is that of the Vocabulary part of C2. C2-V seems to serve as a "suppressor" variable. The hypothesis could be advanced that the abilities measured by this test are antithetical to those evaluated by the first-quarter freshman English course teachers at Peabody. Certainly no generality can be implied from this scant evidence, however, especially when one compares this beta weight with the three-quarter beta weight of .064 (see Table 2) and considers "sampling fluctuations."

The various beta weights indicate the relative contributions of the independent variables to the prediction of GPR. Therefore, the contribution of Cm + Dm is almost one-half that of ACE-L, is approximately four times that of ACE-Q, is about  $1\frac{1}{4}$  times that of C2-S, is approximately  $1\frac{2}{3}$  times that of C2-L, and is almost one-third that of Eng A.

For three quarters, the multiple R between GPR and the seven test scores is .59. (This R corrected for shrinkage becomes .56.) The predictive equation, using the same notation

<sup>6</sup> McNemar (3, pp. 124-125).

<sup>7</sup> See, for example, Thorndike (4, pp. 201 f.).

<sup>8</sup> Thorndike (4, pp. 335-339).

<sup>9</sup> Fisher (1) and Wherry (5) have reported on the amount of bias in a multiple R when it is applied to a new sample and the correlation with a similar criterion computed. In such a situation we may anticipate that the new R will be smaller than the original R. The multiple R of .62, corrected for such shrinkage, is .59. See Kelley's (2) Formula 12:36, p. 474.

Table 3

Intercorrelations and Beta Weights Involved in the Prediction of Three-Quarter Grade Point Ratio from Combined Scores on Freshman Tests (N = 119)

	0 GPR	I Cm+ Dm	II ACE Total	III C2 Total	IV Eng A
I Cm+Dm	.50		.80	.68	.65
II ACE Total	.48			.84	.69
III C2 Total	.47				.62
IV Eng A	.55				
Beta Weights		.220	-.090	.171	.363

as above, is

$$z_0' = .242z_1 + .013z_2 - .115z_3 + .064z_4 \\ - .082z_5 + .166z_6 + .367z_7.$$

Here the contribution of Cm + Dm is about two-thirds that of Eng A and exceeds all other contributions.

It is important to mention that "criterion contamination" may be present in the case of all of the predictors except Cm + Dm. The scores on Cm and Dm were the only test scores which were not made available to the teachers, and therefore they could have had no possible influence on the grading. Cm and Dm contributed substantially under fairly ideal conditions.

In arriving at a third predictive equation, the scores on the tests were combined as shown in Table 3.

The multiple R is .59 (.57, corrected for shrinkage). The regression equation becomes

$$z_0' = .220z_I - .090z_{II} + .171z_{III} + .363z_{IV}.$$

I refers to Cm + Dm, II refers to ACE Total, III to C2 Total, and IV to Eng A.

The use of Cm + Dm and Eng A in a fourth predictive equation, using again three-quarter GPR, indicates that the inclusion of Cm + Dm does not greatly enhance the original correlation between Eng A and GPR, though the multiple R, of course, represents a more stable measure. The multiple R is .57. The regression equation becomes

$$z_0' = .212z_I + .412z_{IV}.$$

where I refers to  $Cm + Dm$  and IV refers to Eng A.

It is interesting to note that in every case, Eng A is represented by the highest validity coefficient and the largest beta weight. The fact that English is a required course for freshmen at Peabody and mathematics is not may help explain the larger predictive value of some of the English tests, as compared with ACE-Q; but it does not contribute much toward an explanation of the differences between Eng A and the other English tests. In considering other factors extrinsic to the test itself which might account for some of the high degree of relationship between Eng A and GPR, the fact that Eng A was administered last comes to mind. The time limit (40 minutes) was adequate for far more than half of the testees to finish Eng A; the same was not true of any of the other tests. Perhaps in addition to knowledge of certain English fundamentals, Eng A tested, for this group, motivation and perseverance to a greater degree than any of the other tests. Perhaps, further, these qualities are closely related to freshman GPR at Peabody.

#### Summary

1. Two easy tests (Otis Quick-Scoring Mental Ability Tests, Beta Tests, for Grades 4-9, Forms Cm and Dm) were administered to a group of entering college freshmen for the purpose of giving the students practice on objective tests and acquainting them with the mechanics of a machine-scored answer sheet.

2. These easy practice tests showed some predictive values above those of several more conventional placement tests; namely, American Council on Education Psychological Examination (1949 College Edition) and Co-operative English Test C2 (Form S). The usefulness of the inclusion of the Otis tests in such a battery is indicated.

3. The Cooperative English Test A: Mechanics of Expression (Form T), in every combination with the other placement tests, contributed most substantially to prediction of freshman GPR.

4. The numerically largest beta weight in the one-quarter regression equation was that of C2-Vocabulary, which in this sample appeared to act as a "suppressor" variable. The corresponding beta weight in the three-quarter predictive equation was small but was positive instead of negative.

Received August 26, 1952.

#### References

1. Fisher, R. A. Influence of rainfall on the yield of wheat at Rothamstead. *Philos. Trans.*, 1923, 213, 89-142.
2. Kelley, T. L. *Fundamentals of statistics*. Cambridge: Harvard Univ. Press, 1947.
3. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
4. Thorndike, R. L. *Personnel selection: test and measurement techniques*. New York: Wiley, 1949.
5. Wherry, R. J. A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Ann. math. Statist.*, 1939, 2, 440-457.

## Development of a Short Test to Predict a Complex Aggregate Score<sup>1</sup>

Helen Tomlinson and John T. Preston

*USAF Training Command, Human Resources Research Center, Personnel Research Laboratory, Lackland Air Force Base, San Antonio, Texas*

In a large organization, job assignment involves administration to potential employees of a wide variety of tests to predict probability of success in different kinds of assignments. This is an expensive procedure, and most organizations which require continuous hiring have worked out means of spotting applicants who have little chance to prove successful in any current job openings. One of these is the short preliminary test with cut-off scores based on empirically determined probabilities of meeting qualifying requirements for job appointment.

This article describes a technique for constructing a short test to determine probability of reaching a qualifying score on an aggregate score empirically weighted for prediction of success in one group of training schools. The method was developed in connection with the Service problem of recruitment to meet manpower needs in a specialized job area.

### Method

In the Air Force problem, the required short test is to predict probability of reaching a qualifying cut-off on a weighted aggregate score, the aptitude index. Each aptitude index combines the variables of the Airman Classification Battery for optimum prediction of training success in one cluster of related technical specialties.

A preliminary form, twice the length specified for the final predictor test, was composed of blocs of items, with each bloc patterned after one of the components of the aptitude index. Analysis of results from a try-out of the preliminary form determined selection of items for the final form. Test scores included in the aptitude index are converted to normalized standard AF scores before weighting them into the aggregate. Sections of the pre-

dictor test are weighted by the number of items so that the total raw score is appropriately weighted. Table 1 shows the composition of the criterion, and the preliminary and final forms of the predictor.

Items in five of the six subtests were selected from appropriate item pools for significant discrimination against the corresponding test of the aptitude index. Experience items were selected and keyed for positive correlation with both the Biographical Inventory score and the aptitude index. Because of the restricted choice of items for Subtest 6, a group of items in a related area, Test VII, was included in the preliminary form for possible use if the items of Subtest 6 did not hold up in the analysis.

The preliminary 64-item form was administered to two unselected samples which had about equal representation of men assigned for basic training to Sheppard and to Lackland Air Force Bases.

For analysis purposes, each sample of 370 was split into the Sheppard group and the Lackland group. Table 2 shows the composition of the samples and score distribution statistics for the preliminary test, the criterion, and a vocabulary test.

Table 1  
Composition of the Criterion, the Preliminary Form, and the Final Form of the Predictor

Test	Criterion		Sub-test	Predictor	
	No. of Items	Weight		No. of Items	
				Prelim.	Final
I	35	2	1	12	6
II	143	2	2	12	6
III	30	1	3	7	3
IV	30	2	4	10	6
V	15	1	5	6	3
VI	30	2	6	10	6
VII	20	-	7	6	-

<sup>1</sup> The views expressed in this article are those of the authors and do not necessarily represent the official views of the United States Air Force.

Table 2

Raw Score Distribution Statistics for the Preliminary Test and Standard AF Score Distribution  
Statistics for the Aptitude Index and a Vocabulary Test

	N	Preliminary Test		Aptitude Index		Vocabulary Test		Correlation of Preliminary Test with Aptitude Index
		M	SD	M	SD	M	SD	
Sample I								
Lackland	194	41.1	11.0	5.5	2.1	6.0	1.8	.86
Sheppard	176	38.2	10.7	5.4	2.1	5.5	1.9	.87
Sample II								
Lackland	193	41.2	10.7	5.6	2.2	5.6	2.0	.86
Sheppard	177	38.3	11.1	5.2	2.1	5.3	2.0	.86

Table 3

Distribution Statistics and Correlations of the Predictor\* with the Criterion

	N	Predictor		Aptitude Index		Correlation of Predictor with Aptitude Index
		M	SD	M	SD	
Sample I						
Lackland	194	19.5	5.8	5.5	2.1	.87
Sheppard	176	17.7	5.9	5.4	2.1	.85
Total	370	18.7	5.9	5.5	2.1	.85
Sample II						
Lackland	193	19.4	6.1	5.6	2.2	.86
Sheppard	177	17.8	6.0	5.2	2.1	.84
Total	370	18.6	6.1	5.4	2.1	.85
Sample III (Independent)	882	17.0	6.6	4.9	2.1	.86

\* Scoring keys for the selected 30 items applied to answer sheets for the preliminary test.

Selection of items for the final form was based on several criteria: (1) Correlation ( $r$ ) with the aptitude index; (2) Correlation ( $r$ ) with the parent test; (3) Lower discrimination for other tests of the aptitude index composite than for the parent test; and (4) Difficulty index centering around 60%  $R$  (uncorrected).

Effect of chance error was diminished by running the item statistics separately for the two samples. Instead of preparing separate 30-item keys for each sample, the criterion of consistency between samples was added for producing a single 30-item key.

#### Efficiency of Prediction

Application of the 30-item key to the experimental answer sheets yielded the distribution statistics and correlations appearing in

Table 3. The key was developed on Samples I and II. Sample III is independent.

Since the new test is essentially a short form of the aptitude index, the predictor has a special relation to the criterion. The Kuder-Richardson estimate of reliability (Formula

Table 4

Correlations with Control Aptitude Index  
(Sample: 324 Lackland Airmen)

	M	SD	Correlation with Control Aptitude Index
Control Aptitude Index	5.8	2.1	
Criterion Aptitude Index	5.7	2.1	.77
Predictor Score	19.9	5.7	.65

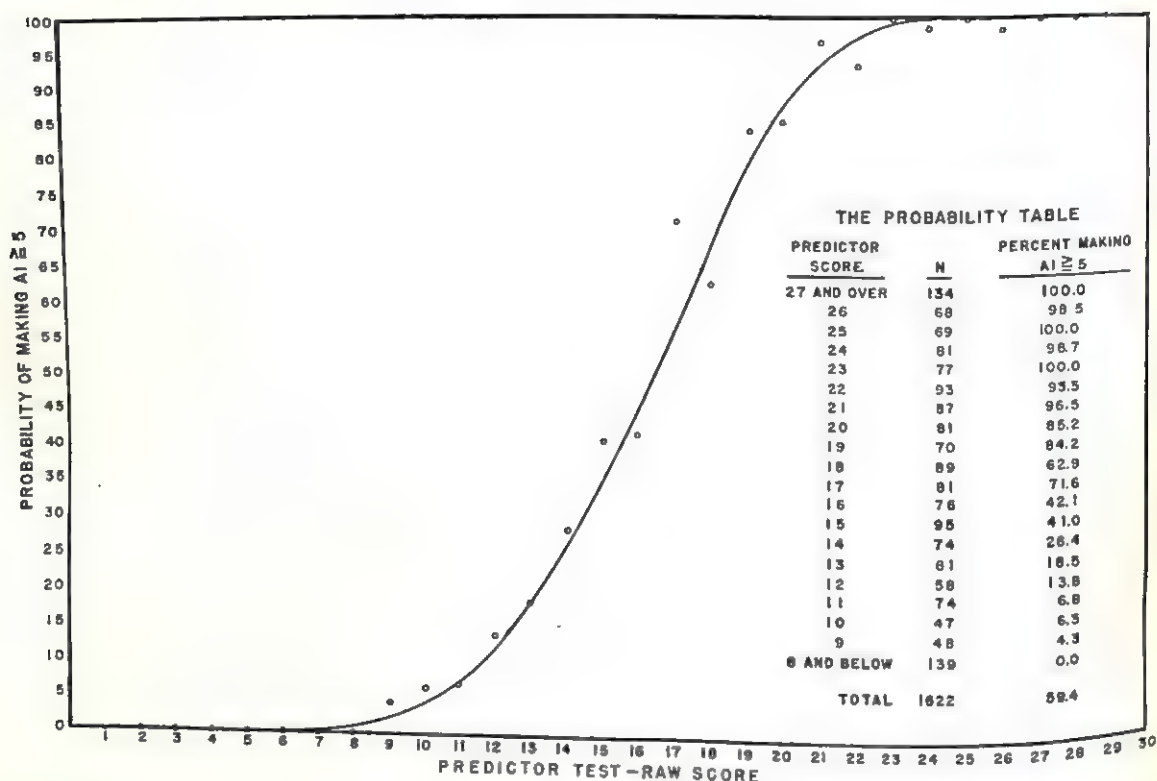


FIG. 1. Curve showing probability of making an aptitude index of 5 or more for each raw score on the predictor test.

20) is .84. The estimate obtained by correlating scores for the final form against scores for the remaining 34 items is .82. These reliability coefficients do not differ significantly from the coefficients of prediction (.84 - .87).

To show that the predictor test is specifically predictive of the criterion, the correlation was computed with a control aptitude index of median correlation with the criterion aptitude index and maximum correlation with the Armed Services Qualification Test. Table 4 shows these correlations.

Because of common components in the eight Air Force aptitude indexes, their intercorrelations are high. Consequently the correlation of the predictor with the control aptitude index (.65) is less than the correlation between the criterion aptitude index and the control aptitude index (.77).

Rather than a direct score conversion table, empirical probabilities for achieving an aptitude index of at least 5 were computed from the combined sample of 1622 for each pre-

dicator score directly from the 30 x 9 regression chart. Figure 1 shows that the curve of percentages closely approximates the expected ogive. The steepness of the curve between predictor scores 16 and 17 suggests a reliable cut-off for a better-than-even chance of qualifying for training in the specific job area.

### Summary

This method of constructing short screening tests offers two advantages:

1. The screening process can be readily directed to current personnel requirements.
2. A simple "Rights" score read into a table of probabilities makes it possible to use clerical assistants with only a minimum of training for administering, scoring, and interpreting the tests.

The method is applicable only in conjunction with reliable, well-validated selection instruments.

Received November 10, 1952.

## The Classification of Occupations by Means of Kuder Interest Profiles: I. The Development of Interest Groups

John L. Holland, Allen H. Krause, M. Eloise Nixon, and Mary F. Trembath

*Vocational Counseling Center, Western Reserve University*

The need for a more extensive knowledge of occupational interests is well recognized by vocational counselors. The empirical evidence is limited largely to the data concerning the Strong Vocational Interest Blank and the Kuder Preference Record. Since these inventories have been applied to only a small number of occupations, it is frequently necessary in practice to speculate about the nature of the interests for many occupations. An "unknown" occupation presents at least two problems: (1) What are the characteristic interests of this occupation?, and (2) In what known interest group does this occupation belong?

These problems are commonly approached by developing interest keys using Strong's method or by securing Strong interest profiles for unknown occupations. Since these methods present practical and financial problems, a simpler, less expensive method of measuring and classifying interests for a large number of occupations would accelerate the extension of our knowledge of interests and interest groups.

This study is an attempt to classify occupations empirically using KPR profiles and to present a research tool for validation. While the sets of interest groups developed here are empirical in nature, their predictive power, especially with respect to job satisfaction, is largely unknown. It is hoped that others will be able to test and extend this classification system by making cross-comparisons with the Strong, and by predicting job satisfaction or vocational choice.

### Method

A group of 45 KPR profiles representing the major interest groups for men, as defined by the SVIB, was selected from the Kuder manual of instructions (4). This selection was made since the Strong groups represent a classification system supported by consider-

able evidence including a number of factor analyses by Strong (6) and Thurstone (8). It is assumed that a new method of classification would reveal comparable interest groups.

About 27 of the 45 occupations used in this sample are similar to, if not identical with, the Strong sample of 44 occupations. Eighteen additional Kuder profiles were used to define more accurately the limits of a given interest group, and to increase the representativeness of the group by employing a larger sample of occupations. This need is especially apparent in the case of groups defined by one to three occupations.

A second group of 42 profiles was secured for women. No attempt was made to select these profiles in terms of occupational groups; however, apprentice and military occupations, except for aviation assembly and repair worker, were avoided. Also, housewives as an occupational group were not used since the Kuder sample is limited to only three groups of housewives (wives of lawyers, physicians, and farmers) and may accordingly be quite unrepresentative of housewives-in-general.

For females, the Kuder sample contains about 17 of the 25 occupations in the present Strong Blank. The twenty-five additional occupations again were used to define the limits and to increase the representativeness of the interest groups.

Using  $\rho$  as an index of profile similarity, the Kuder profiles were intercorrelated.<sup>1</sup> By grouping profiles which intercorrelate .70 or greater, a set of interest groups was derived

<sup>1</sup> Profiles were coded by listing the highest scale first with the remaining scales following in descending order. For these computations the outdoor scale was omitted since it was available for only a few occupations. The matrices formed by the intercorrelation of profiles (16 pages) are deposited with the American Documentation Institute. Order Document 3725 from American Documentation Institute, Library of Congress, Washington 25, D. C., remitting \$1.00 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$2.40 for photocopies (6 X 8 inches) readable without optical aid.

for each sex. This procedure consisted of a simple cluster analysis. Each profile was inspected for its pattern of intercorrelations. Occupational groups were formed by classifying profiles having similar intercorrelational patterns. In most cases the evidence for a particular grouping is clear-cut; however, when a profile correlates with about equal frequency with the occupations of two or more groups, it is listed in each group in order to give an accurate picture of the data.

It seemed meaningful to make several exceptions to this method of classification. In the case of males, the profiles for farmer, aviator, carpenter, and forest supervisor are placed in the interest group with which they show the highest single correlation. For females, this relaxation of the criterion was made for stenographer and typist, and sales

clerk. Although these profiles fail to meet the criterion for inclusion, their intercorrelational patterns are similar to the typical pattern of their respective groups.

In order to secure a useful arrangement of occupations within a given interest group, occupations are ordered in terms of their "representativeness." The occupation which shows the greatest number of correlations equal to or greater than .70 with the other members of the same group is listed first in that group. This occupation is designated as the "core" occupation. The remaining occupations are then arranged in descending order of correlation with the core profile. Ties in rank were resolved by comparing the number of correlations equal to or greater than .70 for each of these occupations with the other occupations in the same group. The relative position of

Table 1  
Kuder Interest Groups (Male)

Group	Occupation	Rho	Code*	N
I. Skilled and Technical				
	Asst. District Rangers		0'1325687'94	28
	Electricians	.95	1'30528697'4	37
	Vocational Training Teachers	.87	'15382697'4	35
	Machinists	.80	'153279864'	117
	Meteorologists** (2)	.80	3'6127589'4	185
	Engineers (3)	.78	'3216547098'	653
	Draftsmen	.77	5'13279684'	216
	Aviators	.52	31'796528'4	34
	Farmers	.39	0'913825647'	129
	Carpenters	.32	'1058973246'	124
II. Managerial				
	Production Managers		'2631495807'	139
	Engineers (3)	.79	'3216547098'	653
	Controllers	.75	23'916475'8	24
III. Scientific				
	Psychologists		36'285791'4	111
	H. S. Teachers of Mathematics	.87	23'867519'4	30
	Laboratory Technicians	.81	3'56278914'	90
	All Physicians and Surgeons	.78	3'658071294'	260
	Chemists	.76	3'672518'49	54
	District Rangers	.73	0'256381749'	102
	Meteorologists (2)	.68	3'6127589'4	185
	Protective Service Occupations	.66	'6870351249'	23
	Engineers (3)	.43	'3216547098'	653
IV. Drugstore Managers and Pharmacists			'342987561'	140

\* Scales to the left of the first apostrophe indicate scales at or above the 75th percentile. Scales to the right of the last apostrophe are at the 25th percentile or below.

\*\* The number in parentheses following occupational title designates occupations listed in two or three interest groups.

Table 1—Continued

Group	Occupation	Rho	Code*	N
V. Welfare				
	H. S. Teachers of Social Studies	.88	8'6243795'1	23
	School Administrators	.83	8'62934751'	65
	Social and Welfare Workers	.68	8'6439257'1	53
	Forest Supervisors	.68	0'682543197'	17
	Clergymen	.63	86'793254'1	43
VI. Clerical				
	General Office Clerks	.98	9'267438501'	110
	General Accountants	.91	29'64730851'	92
	Office Mgrs. and Chief Clerks	.91	2'964758301'	138
	Cost Accountants	.83	29'6475318'	28
	Printers and Pressmen	.78	'976253418'	32
	Financial Institution Clerks	.75	2'98674351'	25
	Lawyers and Judges (2)	.73	6'97428503'1	331
	Purchasing Agents and Buyers (2)		'469278531'	103
VII. Expressive				
	Sales Managers	.98	4'768953201'	230
	Life Insurance Salesmen	.96	4'7869532'1	50
	Salesmen, to Consumers	.90	4'678935012'	353
	Reg. Salespersons, Dept. Store	.90	4'7965823'1	184
	Stock and Bond Salesmen	.88	4'8796523'1	118
	Advertising Agents	.87	64'75893'12	26
	Route Salesmen	.85	4'87569312'	104
	Authors, Editors, and Reporters	.83	6'794850'321	113
	Personnel Managers	.82	'486792305'1	50
	Banking, Finance, and Insurance Officials	.75	'4796258031'	42
	Lawyers and Judges (2)	.70	6'97428503'1	331
	Purchasing Agents and Buyers (2)	.70	'469278531'	103
	Musicians and Music Teachers	.70	7'65894'321	77
	Commercial Artists	.50	5'6794281'3	31

an occupation within its group serves then as a crude index of its communality or belongingness.

Each set of interest groups was arranged by inspecting the matrix formed by the core occupations. The highest negative correlation between a pair of core occupations designates the first and last interest groups, or the groups whose interests are most divergent. The remaining groups are placed between these groups in accordance with their correlation with the first group listed.<sup>2</sup>

### Results

**Males.** The classification of profiles produced seven interest groups which have been

<sup>2</sup> If the remaining groups are arranged in accordance with their correlation with the last group, groups V and VI will change position, but the original arrangement is essentially the same.

designated tentatively as: I. Skilled and Technical; II. Managerial; III. Scientific; IV. Drugstore Managers and Pharmacists; V. Welfare; VI. Clerical; and VII. Expressive. Table 1 shows the male sample classified into these groups. To increase the usefulness of Table 1, the *rho* with the core occupation, the *N*, and the code<sup>3</sup> for each occupation have been listed.

In general, the obtained groups are similar to related Strong Groups; the Kuder Groups I, II, III, V, and VI agree in content with the Strong Groups IV, III, I, V, and VIII, respectively. The differences between comparable groups seem slight.

<sup>3</sup> The attention of readers is directed to the misleading character of a mechanical coding device for counseling purposes that isolates percentiles of 75 or greater or percentiles of 25 or lower. See Diamond, S. The interpretation of interest profiles. *J. appl. Psychol.*, 1948, 32, 512-520.—Editor.

Table 2  
Kuder Interest Groups (Female)

Group	Occupation	Rho	Code*	N
I. Computational				
	H. S. Teachers of Mathematics		2'31597864'	47
	Office Machine Operators	.93	2'13598746'	62
	Tearoom and Restaurant Mgrs.	.73	2'3514697'8	20
	Hospital Dietitians** (2)	.70	3'21765849'	31
	Statistical Clerks	.43	2'36574189'	27
II. Scientific and Technical				
	H. S. Teachers of Home Economics		'513867249'	136
	Aviation Assembly & Repair Wrks.	.90	1'3578624'9	75
	Cooks and Bakers	.88	'16347582'9	31
	Physicians	.87	31'85627'49	43
	All Trained Nurses	.83	8'35176249'	1071
	Supervisors and Head Nurses	.70	'831752649'	196
	Dental Hygienists	.70	83'571462'9	35
	Laboratory Technicians	.70	31'527486'9	31
	Occupational Therapists	.63	15'738624'9	70
	Home Demonstration Agents	.62	'85147326'9	24
	Hospital Dietitians (2)	.43	3'21765849'	31
III. Clerical				
	H. S. Teachers of Commercial Subjects		2'94756138'	64
	General Office Clerks	.85	'295471368'	136
	Bookkeepers	.80	2'96754381'	62
	Stenographers and Typists	.32	'973562148'	235
IV. Linguistic				
	Journalists		6'5473182'9	31
	Copy Writers, Mail Order Co.	.95	64'57132'89	19
	Librarians	.78	6'57149328'	39
	Artists and Art Teachers	.75	5'167483'29	22
	H. S. Teachers of Language (2)	.75	6'75489132'	42
	H. S. Teachers of English (2)	.71	6'75482193'	110
V. Expressive				
	Assistant Buyers, Dept. Store		46'7852193'	58
	Personnel Mgrs., Mail Order Co.	.87	4'68751932'	34
	H. S. Teachers of English (2)	.86	6'75482193'	110
	H. S. Teachers of Social Studies	.80	'684527193'	56
	Social Workers	.77	8'647531'29	50
	H. S. Teachers of Language (2)	.75	6'75489132'	42
	Personnel Wrks. other than Mgrs.	.73	48'6217953'	27
	Musicians and Music Teachers	.70	7'5684192'3	68
	Cashiers	.70	74'6529381'	43
	Retail Buyers	.70	4'7625139'8	29
	Salespersons, Dept. Store, Reg.	.70	4'75896231'	617
	Religious Workers	.68	8'6754139'2	31
	Floor and Section Mgrs., Dept. St.	.65	4'68297135'	25
	Secretaries	.65	'467952138'	121
	Executives, Mail Order Company	.56	4'27653981'	67
	Teachers, Primary and Kindergarten	.50	'857621493'	544
	Office Mgrs., Chief Clerks	.40	'426179358'	29
	Sales Clerks	.38	'849761325'	26
	Telephone Operators	.22	'495781236'	22

\* Scales to the left of the first apostrophe indicate scales at or above the 75th percentile. Scales to the right of the last apostrophe are at the 25th percentile or below.

\*\* The number in parentheses following occupational title designates occupations listed in two or three interest groups.

In the Kuder classification, engineers are classified in the skilled, scientific, and managerial groups rather than in science alone as in the SVIB. District rangers are classified in both the skilled and scientific groups rather than in science alone. Protective service occupations are classified in science rather than in skilled trades. Personnel managers are classified in the expressive group rather than social service. Lawyers are classified in the clerical and expressive groups rather than language occupations only. Printers and pressmen are classified in the clerical group rather than in the skilled trades. Purchasing agents and buyers are classified in both the clerical and expressive groups rather than in business detail alone. Pharmacists and drugstore managers form a separate group in the Kuder data while they fall in Group VIII on the Strong.

These differences are due in part to the classifying of occupations in one or more groups by the KPR, and to the limiting of an occupation to a single group by the SVIB. Consequently, only four of the above occupations are placed in opposed interest groups: protective service occupations, personnel managers, printers and pressmen, and pharmacists and drugstore managers.

The creation of Group VII, Expressive Occupations, presents the greatest difference between these interest systems. Group VII incorporates Strong Groups IX and X and a number of other occupations: salesmen to consumers; regular salespersons (department store); stock and bond salesmen; route salesmen; personnel managers; banking, finance, insurance officials; purchasing agents and buyers; musicians and music teachers; and commercial artists. The cohesiveness of Group VII is marked. The first four occupations intercorrelate .70 or greater with each other. Furthermore, the median  $r_{ho}$  for the entire group of 14 occupations with the core occupation, sales manager, is .86.

While it is possible to fractionate Group VII into three sub-groups consisting of sales, language, and art (including music), there is no empirical justification for such groups since the correlational patterns show no sharp

differences. This procedure would, of course, make the two interest systems appear more alike.

*Females.* The correlational matrix for females produced five interest groups: I. Computational; II. Scientific and Technical; III. Clerical; IV. Linguistic; and V. Expressive. The comparability of these groups with the 13 groups shown by Strong (7) is difficult to ascertain. The Strong groups are defined frequently by a single occupation so that their limits are not clearly structured.

Kuder Group I, Computational, includes Strong Group XII (mathematics and science teacher) and three additional occupations not found in the Strong sample: office machine operators, tearoom and restaurant managers, and statistical clerks. Hospital dietitian in Strong Group IX falls in this group and also in Group II, Scientific and Technical.

Kuder Group II, Scientific and Technical, includes Strong Groups IX, XI, and XIII (home economics teacher, dietitian, occupational therapist, nurse, laboratory technician, and physician) as well as a number of occupations not contained in the SVIB.

Kuder Group III, Clerical, is a rough approximation of Strong Group VII as both groups include high school teachers of commercial subjects, general office workers, bookkeepers, and stenographers.

Kuder Group IV, Linguistic, combines Strong Groups I and II (artist, librarian, and English teacher) and a number of language occupations not included in the SVIB.

Kuder Group V, Expressive, combines Strong Groups III, IV, VI and VIII (social worker, social science teacher, elementary school teacher, and buyer) as well as a number of related sales, social service, and administrative occupations. High school teachers of English which occur in Kuder Group IV are also listed in Group V.

In general, the Kuder set of interest groups appears to be a coarser classification system than the Strong set, but the direction of their classification appears essentially the same. A more accurate picture of the relationships between these systems can be obtained by a comparison of their correlational matrices.

Discussion

The occupational classification system presented here is limited by the original data. The representativeness of most of the occupational profiles is largely unknown; furthermore, each criterion group contains both successful and unsuccessful, experienced and inexperienced, as well as satisfied and dissatisfied workers (4). The usefulness of this interest system in interpreting individual profiles is consequently restricted.

The validity of these interest groups is problematical. Tests of significance are inappropriate because of the large number of computations and because of the unorthodox use of *rho*; that is, varying N's and yet only eight degrees of freedom for each correlation. The use of *rho* as an index of profile similarity violates the assumptions underlying its application. Moreover, as Guilford (3) points out, the "ipsative property of the Kuder scores . . . renders their use for intercorrelations among themselves . . . so questionable as to preclude attempts at analysis by the R-technique."

These difficulties pose a choice of abandoning the evidence until more adequate data can be obtained and statistical elegance can be achieved, or to test the interest groups by means of a prediction study. The latter alternative seems desirable in view of the need for an immediate understanding of the present data. Although a predictive test may make statisticians wince, it will yield an estimate of the predictive power of the classification.

While they furnish only suggestive and incomplete evidence, several related studies support the Kuder interest groups. In a factor analysis of the SVIB, KPR, Bell Inventory, and MMPI, Cottle (2) extracted seven factors similar to those described by Thurstone (8) and Strong (6). Cottle's interest factors support the Kuder classification including Group VII, Expressive, which combines Strong Groups IX and X in addition to a number of other occupations. The factor G obtained by Cottle appears similar to the Kuder Group VII since it is characterized by high positive loadings on Strong Groups IX and X and high negative loadings on the Kuder mechanical scale. Similarly, an in-

spection of the present Kuder matrix indicates that the expressive occupations correlate most negatively with the skilled trade occupations.

In addition, Cottle's matrix for the SVIB with the KPR reveals that the Strong Group keys correlate with the KPR scales in a pattern which is similar to the coded profile for the comparable Kuder interest group. The positive and negative correlations of Kuder scales with a given Strong group scale produce coded profiles which are similar to those in the Kuder Groups. For example, the code for high school teachers of social studies, the core occupation for Kuder Group V, Welfare, is 862437951. Cottle's matrix produces the code 867495213. The other five group scales show similar relationships with the Kuder interest groups.

In a study of occupational test patterns, Barnette (1) has supplied Kuder profiles for "success" and "failure" samples for five occupational groups: engineers, accountants, clerical personnel (except accountants), a specialized clerical group (primarily verbal in nature), and salesmen. When the pairs of "success" and "failure" profiles are correlated with comparable Kuder core occupations, six of the seven tests reveal that the "success" groups correlate higher with the core occupation than do the "failure" groups. The data

Table 3  
The Relation of Occupational Success and Failure to Core Occupations of Corresponding Kuder Interest Groups\*

Occupations		N	I	II	III	VI	VII
Engineers	S	83	.83	.66	.44		
	F	39	.59	.21	-.09		
Accountants	S	74					.92
	F	24					.85
Clerical (Specialized)	S	20					.78
	F	14					.47
Clerical (General)	S	54					.83
	F	40					.48
Salesmen	S	77					.86
	F	56					.94

\* Correlations are *rho* coefficients.

suggest that the degree of similarity of an individual profile to a given interest group is related to success or failure in the occupation. Table 3 shows these relationships.

The classification of occupations by means of the entire profile appears to have several advantages which other systems such as the method proposed by Weiner (12) lack. Systems of classification which rely on the two or three highest scales, or the scales above a certain cutting score, utilize only a portion of the data. There is no experimental evidence revealing that "high" scores are more significant or useful than "low" scores. In practice, a coding system which uses both low and high scores impels the counselor to notice and use more of the data. Finally, a total code minimizes the misclassification of occupations. If only the high scores are used to categorize an occupation, occupations which have similar patterns may be sorted into different groups by means of one or two scales which may represent the only differences between a pair of profiles.

The results suggest further that previous studies of the SVIB and KPR (5, 9, 10, 11, 13) may be in error in concluding that these instruments show little relationship. Their findings may be largely negative due to an inappropriate level of analysis; that is, a single scale comparison rather than a pattern method.

#### Summary

A sample of KPR profiles was intercorrelated and classified by means of a simple cluster analysis. Sets of occupational groups for men and women were derived which are comparable with many of the SVIB groups. While these groups are empirical in nature, their validity is unknown. It was suggested that validation might be established by pre-

diction studies of job satisfaction and vocational choice.

Received August 29, 1952.

#### References

1. Barnette, W. L. Occupational aptitude patterns of selected groups of veterans. *Psychol. Monogr.*, 1951, 65, No. 5 (Whole No. 322).
2. Cottle, W. C. A factorial study of the Multi-phasic, Strong, Kuder, and Bell inventories using a population of adult males. *Psychometrika*, 1950, 15, 25-47.
3. Guilford, J. P. When not to factor analyze. *Psychol. Bull.*, 1952, 49, 26-37.
4. Kuder, G. F. *Examiner manual for the Kuder Preference Record, vocational, form C, second revision*. Chicago: Science Research Associates, March 1951.
5. Peters, E. F. Vocational interests as measured by the Strong and Kuder inventories. *Sch. and Soc.*, 1942, 55, 453-455.
6. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford Univ. Press, 1943.
7. Strong, E. K., Jr. *Manual for vocational interest blank for women*. Stanford: Stanford Univ. Press, 1951.
8. Thurstone, L. L. A multiple factor study of vocational interests. *Personnel J.*, 1931, 10, 198-205.
9. Triggs, F. O. A study of the relation of Kuder Preference Record scores to various other measures. *Educ. Psychol. Measmt.*, 1943, 3, 341-354.
10. Triggs, F. O. A further comparison of interest measurement by the Kuder Preference Record and the Strong Vocational Interest Blank for men. *J. educ. Res.*, 1944, 37, 538-544.
11. Triggs, F. O. A further comparison of interest measurement by the Kuder Preference Record and the Strong Vocational Interest Blank for women. *J. educ. Res.*, 1944, 38, 193-200.
12. Wiener, D. N. Empirical occupational groupings of Kuder Preference Record profiles. *Educ. Psychol. Measmt.*, 1951, 11, 273-279.
13. Wittenborn, J. R., Triggs, F. O. and Feder, D. D. A comparison of interest measurement by the Kuder Preference Record and the Strong Vocational Interest Blanks for men and women. *Educ. Psychol. Measmt.*, 1943, 3, 239-257.

## The Validity of the Mooney Problem Check List \*

Charles J. McIntyre

*The Pennsylvania State College*

The Problem Check List is an instrument developed by Mooney (1) to enable the teacher or counselor to quickly identify problems or problem areas which concern his students. The high school form, which was used in this study, consists of 330 problems found to be of particular concern to students. They are classified into the following eleven major areas: (1) Health and Physical Development; (2) Finances, Living Conditions, and Employment; (3) Social and Recreational Activities; (4) Courtship, Sex, and Marriage; (5) Social-Psychological Relations; (6) Personal-Psychological Relations; (7) Morals and Religion; (8) Home and Family; (9) The Future: Vocational and Educational; (10) Adjustment to School Work; and (11) Curriculum and Teaching Procedures.

Normally the subject is instructed to underline those problems that bother him and to circle those underlined problems which trouble him the most. In this study no distinction was made between underlined and circled items.

Mooney (2) has said that the nature of the Check List makes it impossible to arrive at a definitive conclusion about its validity. Validity, he says, must be determined in terms of the particular purpose and the particular situation. While it probably is true that conventional measures of validity are difficult if not impossible to obtain for an instrument of this kind, it appears nevertheless that the Check List should meet at least three minimum requirements: (1) Students recognize their own problems; (2) They find these problems listed on the Check List; and (3) They are willing to record them.

This study assumes that if these three conditions are met it should be possible to predict the relative number of problems listed by particular groups of students in particular

areas. Hence the following hypotheses were formulated: (1) The less intelligent students would have more problems than the more intelligent in the area of Adjustment to School Work; (2) Seniors would have more problems than those in the lower grades in the area of The Future: Vocational and Educational; (3) Students from broken homes would have more problems than those from intact homes in the area of Home and Family; (4) Boys would have more problems than girls in the area of Adjustment to School Work; (5) Boys would have more problems than girls in the area of The Future: Vocational and Educational; (6) Negroes would have more problems than whites in the area of Finances, Living Conditions, and Employment; and (7) Girls would have more problems than boys in the area of Courtship, Sex, and Marriage.

The rationale behind each of the hypotheses should be evident.

### Procedure

*Subjects.* The subjects were 407 high school students in grades ten to twelve inclusive. The school which they attended was the only public high school in a highly industrial Pennsylvania city with a population of approximately sixty thousand. The city population is highly heterogeneous in terms of race, religion and national origins, and this heterogeneity is reflected in the school population.

*Method.* Approximately one-fourth of the school population was sampled. The Check Lists were administered by the homeroom teachers during the period prior to the first class in the morning. Homerooms were selected so that the proportion of students in the several courses and classes in the sample would approximate the proportion of students in these courses and classes in the entire school population. In this way it was possible to secure a reasonably representative sample

\* This paper represents the substance of a thesis submitted in partial fulfillment of the requirements for the M.S. degree at The Pennsylvania State College.

Table 1

Groups and Problem Areas Relevant to Hypotheses 1 to 6 with N, Mean, SD and CR for Each

Hypothesis	Problem Area	Group	N	Mean	SD	CR
1.	Adjustment to School Work	Less intelligent	55	5.1	.28	3.40
		More intelligent	61	2.9	.12	
2.	The Future: Vocational and Educational	Seniors	156	3.5	.07	3.39
		Sophomores	157	2.4	.04	
3.	Home and Family	Broken Home	85	2.8	.10	2.49
		Intact Home	318	1.9	.02	
4.	Adjustment to School Work	Boys	202	4.6	.07	2.28
		Girls	204	3.8	.05	
5.	The Future: Vocational and Educational	Boys	202	3.3	.05	2.14
		Girls	204	2.6	.04	
6.	Finances, Living Conditions and Employment	Negro	100	3.5	.08	2.11
		White	295	2.8	.03	

while retaining the administrative convenience of intact homerooms.

In questionnaires of this kind the problem of a student's honesty is a serious one, particularly when there is a chance, as here, that his teachers may check his responses. As an example Olson (3), using the Woodworth-Mathews Personal Data Sheet, found that more symptoms were reported when the questionnaire was left unsigned. Therefore in this study a supplementary instruction sheet was attached to the Check List explaining that the study was being conducted to gather information on the problems of high school students and instructing them not to sign the Check List. They were, however, to put their names on the instruction sheet and hand this in for an attendance record. A system of indiscrete pinholes pricked through both the instruction sheet and the Check List made it possible to later match the two and identify the Check List.

Information on the relevant variables was abstracted from the students' records in the school file. Only two of these variables require further definition: (1) "Less intelligent students" are defined in this study as those students whose Otis Gamma IQ's were more than one standard deviation below the mean IQ of the sample. "More intelligent students" are those whose IQ's were more than one standard deviation above the mean of the

sample; and (2) A student was classified as coming from a broken home if the records indicated that he was not presently living with both natural parents.

*Treatment of the Data.* The mean number of problems reported was computed for each of the variables and problem areas which was relevant with respect to the hypotheses. The hypotheses were tested by computing the critical ratio of the difference between these means.

### Results

1. Hypotheses 1 and 2 were confirmed. In Table 1 it will be seen that the differences between the mean number of problems reported in each case were significant at or beyond the .01 level of confidence.

2. Hypotheses 3, 4, 5, and 6 were confirmed. In Table 1 it will be seen that the differences between means were significant at the .05 level of confidence.

3. Hypothesis 7 was not confirmed. No statistical difference between means was found.

### Summary and Conclusions

The problem of determining the validity of the high school form of the Mooney Problem Check List was attacked by computing the mean number of problems checked in particular problem areas by a group of high

school students who were classifiable into various discrete groups.

This study was founded upon the assumption that the essential test of the validity of an instrument of this kind consists in determining whether or not the students can recognize their own problems, find these problems represented on the Check List, and record them. If these three criteria are met the mean number of problems checked in particular areas by various groups should differ significantly in a reasonable and predictable way.

Hence, seven such differences were hypothesized on rational grounds. That is, because of the sociological and psychological characteristics of particular groups, it was predicted that some groups would check more problems in certain areas than other groups,

providing the three criteria of validity specified above were met by the Check List. Of the seven differences hypothesized, six were found.

It is concluded that these findings present prima facie evidence for the validity of the Check List.

*Received September 2, 1952.*

#### References

1. Mooney, R. L. Exploratory research on students' problems. *J. educ. Res.*, 1943, 37, 218-224.
2. Mooney, R. L. and Price, Mary. *Manual to Accompany Mooney's Problem Check List—High School Form*. Columbus, Ohio: Ohio State University, 1948.
3. Olson, W. C. The waiver of signatures in personal data reports. *J. appl. Psychol.*, 1936, 20, 442-450.

# A Comparison of Manual and College Norms for the MMPI

Fred. T. Tyler and John U. Michaelis

*School of Education, University of California, Berkeley, California*

Various investigators have suggested that the distributions of scores of college students on certain of the MMPI scales are different from those reported in the manual of directions. For instance, McKinley (2) reported that college students obtained higher K-scores than did a less selected (intellectually and educationally) population. Similarly, college men have been found to score high on the masculinity-femininity scale (3), indicating a "deviation of the basic interest pattern in the direction of the opposite sex" (1, p. 5). The MMPI has been extensively used in research and clinical investigations at the college level, so that it should be of interest to know something of the extent to which T-scores obtained from the standardization group correspond to those based upon a college population.

The MMPI (Booklet—short form) was administered to nearly one thousand juniors, seniors and first-year graduate students in education courses in the School of Education at the University of California, either in Education 110 (Educational Psychology) or in those courses required in the programs leading to the General Elementary and General

Secondary Teaching Credentials. T-scores were computed for each of the nine clinical scales<sup>1</sup> for men (N = 470) and women (N = 571) separately. These samples are considerably larger (about 60 and 40 per cent respectively) than those of the original standardization groups. It is recognized that these norms might not be representative of the general college population, since the majority of the subjects were expecting to enter the teaching profession. However, the college norms to be discussed here may be of interest to those who are administering this Inventory to college students, and especially to those concerned with personnel problems in a teacher education program.

It is the purpose of this note to present a brief comparison between the norms reported by Hathaway and McKinley (1) and by Tyler and Michaelis (6). Comparative data are presented in Table 1 for women and Table 2 for men.

For the women, there seem to be some differences between the norms on three scales,

<sup>1</sup> The scores were not corrected by means of the K-score (2, 7).

Table 1  
Raw Score Equivalents of Selected Standard Score Values Based on Manual and U. C. Norms for Female Subjects

Standard Scores												
Scales	80		70		60		50		40		30	
	M*	C†	M	C	M	C	M	C	M	C	M	C
Hs	23	14	17	11	12	7	7	4	1	1	9	10
D	35	33	29	29	24	24	19	19	14	14		
Hy	36	34	30	30	24	26	19	21	13	17		
Pd	26	26	22	22	18	18	13	14	9	10	5	6
Mf	22	24	27	29	31	33	36	37	41	42	46	46
Pa	18	17	15	14	11	12	8	9	5	6	1	4
Pt	36	30	28	24	21	17	13	10	5	4		
Sc	35	28	27	21	19	15	11	9	3	3		
Ma	27	27	23	23	19	19	14	15	9	12	5	8

\* Raw score equivalents based on norms in the Manual (1).

† Raw score equivalents based on college norms (6).

Table 2  
Raw Score Equivalents of Selected Standard Score Values Based on Manual and U. C. Norms for Male Subjects

Scales	Standard Scores											
	80		70		60		50		40		30	
	M*	C*	M	C	M	C	M	C	M	C	M	C
Hs	18	14	13	10	9	7	4	4	0	0		
D	29	34	25	28	21	23	17	18	12	13		
Hy	33	32	27	28	22	24	16	20	11	16		
Pd	26	27	22	23	18	19	14	15	10	10	6	6
Mf	36	45	30	39	25	33	20	27	15	21	10	15
Pa	18	17	15	14	11	12	8	9	5	6	1	4
Pt	31	30	24	23	17	16	10	9	2	1		
Sc	32	30	24	23	17	15	9	8	2	1		
Ma	27	28	23	24	18	20	14	16	9	12	5	8

\* See footnote to Table 1.

Hs, Pt and Sc. Using the manual norms as the basis for interpretation, it is suggested that the college women were less concerned about their health, were freer from fears and lack of confidence, and from bizarre or unusual thoughts (1, 4). On the remainder of the scales there was a high degree of similarity between the norms obtained from the two samples.

Several comments about the two sets of norms are in order:

1. The college men obtained relatively lower scores than did the standardization sample on

only one scale, Hs. Accepting the usual interpretations of this scale, it appears that the college men were less concerned than were these members of the standardization group about the state of their health.

2. The college men appeared to be somewhat less depressed and more feminine in their interests when compared with the members of the original standardization group. For instance, on the D scale, a raw score of 29 has a T-score equivalent of 80 on the manual norms and of 71 on the college norms. Similarly on the Mf scale, a raw score of 30

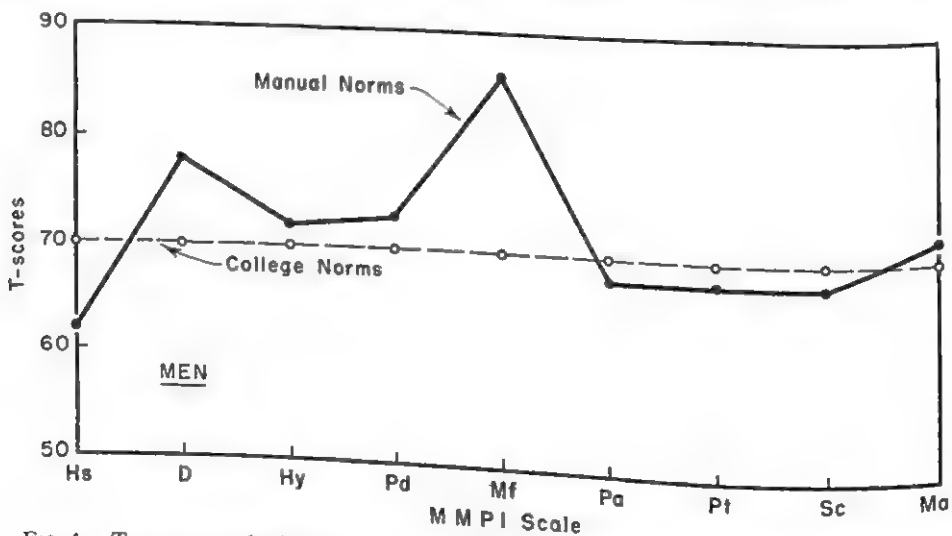


FIG. 1. T-scores on the Manual norms of raw scores corresponding to T-scores of 70 on the College norms (Men).

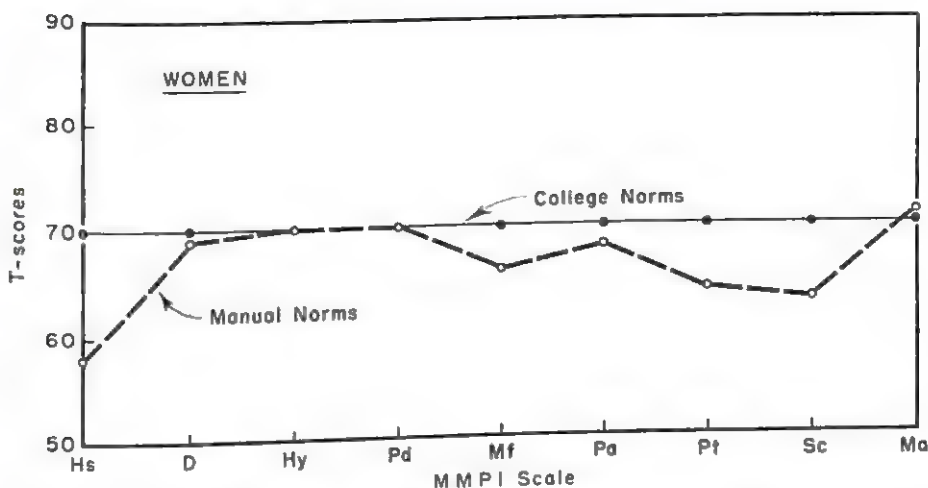


Fig. 2. T-scores on the Manual norms of raw scores corresponding to T-scores of 70 on the College norms (Women).

has T-values of 69 and 55 on the manual and college norms respectively.

3. Raw scores on the remaining scales had very similar T-values on both sets of norms, especially at that part of the scale, the upper end, which is claimed to have clinical significance.

Another method of comparing the two sets of norms was suggested. Raw scores corresponding to T-scores of 70 on the college norms were selected as a basis of comparison. The T-score equivalents on the manual norms were obtained for these same raw scores and are graphed in Figures 1 and 2. A raw score of 70 was chosen because Hathaway and McKinley consider it to be "a borderline score, although useful interpretation will always depend upon the clinician's experience with a given group" (1, p. 8).

At the upper ends of the scales, it is apparent that discrepancies between college and manual norms are found on the Hs, D and Mf scales for men, and on Hs, Pt and Sc for women.

In general, the differences between the two sets of norms appear to be relatively minor.

However, multivariate analysis by means of Hotelling's T (5) might reveal significant differences between the standardization and college groups.

Received October 29, 1952.

#### References

1. Hathaway, S. R. and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation, 1943.
2. McKinley, J. C., Hathaway, S. R., and Meehl, P. E. The Minnesota Multiphasic Personality Inventory: VI. The K Scale. *J. cons. Psychol.*, 1948, 12, 20-31.
3. Nance, R. D. Masculinity-femininity in prospective teachers. *J. educ. Res.*, 1949, 42, 658-666.
4. Tyler, F. T. A factorial analysis of fifteen MMPI scales. *J. cons. Psychol.*, 1951, 15, 451-456.
5. Tyler, F. T. Some examples of multivariate analysis in educational and psychological research. *Psychometrika*, 1952, 17, 289-296.
6. Tyler, F. T. and Michaelis, J. U. *University-student norms for the Minnesota Multiphasic Personality Inventory*. Mimeographed. Available on request to the writers.
7. Tyler, F. T. and Michaelis, J. U. K-scores applied to MMPI scales for college women. (To appear in *Educ. psychol. Measmt.*)

## Socio-economic Status and Culturally-weighted Test Scores of Negro Subjects

Frank C. J. McGurk

*Lehigh University*

A previous article (6) reported on the non-cultural and cultural test scores of 213 pairs of white and Negro high school seniors who had been matched for age, school attendance, school curriculum, and eleven selected socio-economic factors. All subjects were between the ages of sixteen and twenty; the mean age for the whites was 18.1, and the mean age for the Negroes was 18.2. The subjects were obtained from schools in southeastern Pennsylvania and northern New Jersey.

Socio-economic status was defined in terms of the score obtained on a revision of the Sims Record Card. The high Negro socio-economic group is composed of those Negro subjects whose revised Sims scores were in the highest 25% of the range of Negro Sims scores. The low Negro socio-economic group is composed of those Negro subjects whose revised Sims scores were in the lowest 25% of the range of Negro Sims scores.

Test questions were defined as non-cultural and cultural according to the pooled judgments of 78 school teachers, psychologists, and sociologists. Non-cultural questions are those which the judges considered least culturally-weighted; cultural questions are those considered heavily weighted with cultural experiences.

Complete details on the selection and matching of subjects, the revision of the Sims Record Card and the determination of socio-economic status, and the dichotomizing of the questions can be found in the previous article (6).

This paper is concerned with the problem of how the socio-economic status of Negro subjects affects their differential test performance on non-culturally-weighted and culturally-weighted test questions.

### Results

Table 1 shows the mean non-cultural and cultural scores of the high and low socio-economic groups.

Since the standard deviations of the non-cultural and cultural scores are almost identical (4.62 for the non-cultural and 4.64 for the cultural) the findings will be presented in terms of raw scores.

The difference in mean non-cultural score that is associated with a difference in socio-economic level (the H - L difference) is 2.22; the H - L difference for the cultural questions is 1.21. Thus, in the comparison between the

Table 1  
Negro Non-cultural and Cultural Mean Raw Scores by Socio-economic Groups

Type of Question	Socio-economic Group*		Difference (H-L)
	High	Low	
Non-cultural	13.62	11.40	2.22
Cultural	9.81	8.60	1.21
Net change			1.01
SE of net change			0.89
t			1.24
P			20% approx.

\* N for each group = 53.

highest and lowest socio-economic group of Negroes, a greater difference is obtained on the non-cultural, *not on the cultural*, questions. The difference between the two H - L differences is significant only at the 20% level.

### Discussion

In his study of the effects of length of residence in New York City on Negro test scores, Klineberg found that, as the length of residence increased, test scores increased. He also found that his results "... are much clearer for the linguistic tests than for the performance tests" (5, p. 44).

Various writers have described linguistic or verbal tests as sensitive to differences in socio-

economic status because of the culturally-weighted content of such tests (2, 3, 4). Others have described performance or non-verbal tests as not being as sensitive to socio-economic level because the content of these tests is not so culturally-weighted (1, 3, 5). Hence, Klineberg's findings have been interpreted to mean that increasing the socio-economic status of the Negro should be accompanied by a greater improvement on culturally-weighted material than on the less culturally-weighted material.

The present findings do not support such an interpretation. The present data show that the test superiority of the Negro of high socio-economic status over the Negro of low socio-economic status is associated more with a superior performance on the non-cultural questions than on the cultural questions.

Received October 6, 1952.

## References

1. Alpers, T. G. and Boring, E. G. Intelligence test scores of northern and southern white and Negro recruits in 1918. *J. abnorm. soc. Psychol.*, 1944, 39, 471-474.
2. Bean, K. L. Negro responses to certain intelligence test items. *J. Psychol.*, 1941, 12, 191-198.
3. Bean, K. L. Negro responses to verbal and non-verbal test material. *J. Psychol.*, 1942, 13, 343-353.
4. Brown, F. An experimental and critical study of the intelligence of Negro and white kindergarten children. *J. genet. Psychol.*, 1944, 65, 161-175.
5. Klineberg, O. Tests of Negro intelligence. In Otto Klineberg (Ed.), *Characteristics of the American Negro*. New York: Harper & Bros., 1944.
6. McGurk, F. C. J. *Comparison of the performance of Negro and white high school seniors on cultural and non-cultural psychological test questions*. Washington: The Catholic University of America Press, 1951 (microcard).

## The Relationship Between Rater Characteristics and Validity of Ratings

Dorothy E. Schneider and A. G. Bayroff<sup>1</sup>

*Personnel Research Section, Personnel Research and Procedures Branch, TAGO,  
Department of the Army, Washington, D. C.*

The Personnel Research Section, Personnel Research and Procedures Branch, TAGO has been conducting a series of studies aimed at identifying personal characteristics of raters which are associated with more valid ratings. The problem is of particular significance to this office which is responsible for the development of efficiency reports for Army-wide use. It is also important because the absence of more objective criteria makes it necessary to use multiple ratings in the validation of other types of instruments. If personal characteristics of the rater which are related to the validity of ratings he gives could be found, it might be possible to control them, perhaps in the selection of raters, or in the form of a "correction score." While this may not be feasible on official efficiency ratings, it could profitably be done when obtaining ratings for use as criteria. Further, it may be possible to develop new techniques which would be more independent of the personal characteristics of the rater.

### Problem

The study reported here was concerned with the validity of ratings by raters differing in three characteristics: aptitude test score, academic achievement, and rated over-all value to the Army.

### Method

*Subjects.* The population consisted of 400 officers (primarily majors and lieutenant colonels) enrolled as students at the Army Command and General Staff College. The objective of this college is to train potential division commanders and general staff officers, and its students represent a highly selected group. The course was 42 weeks long and

the students were in close contact with each other during the entire period.

The officers were assembled in groups of 33-40, the size of the classes, and asked to rate their class associates. In this manner each officer served as both rater and ratee.

*Design of the Study.* This report covers one aspect of a larger research program on rating methodology. Only those aspects of the design pertinent to the present study will be presented here.

*The Criterion.* The criterion measure was an appraisal of over-all value to the Army. The total population was randomly split into two groups, A and B, each consisting of 200 officers. Officers in Group A provided the criterion rankings for all officers in the population. Each member of this group was required to evaluate his class associates by placing them in the highest, middle, or lowest third of the class. Each rater evaluated 20 ratees from Group A and 20 ratees from Group B; thus each member of the total population was evaluated by 20 raters. The criterion score was the mean of the 20 rankings.

*Anonymous Rating Scale.* This was an 8-point rating scale on which the officers were rated on over-all value to the Army. Each of the scale units was described in some detail. Group A completed 20 such ratings, 10 on officers in Group A, and 10 on officers in Group B. Each of the 400 officers thus received 10 ratings with this instrument, and the mean of these ratings was used as the score. The raters were informed that these were to be held in confidence, that the ratings would not be made available to the College or the Army, and that the raters need not sign their names. These ratings will be referred to hereafter as the unsigned ratings.

*Identified Official Rating Scale.* This was the same 8-point rating scale just described. It was used by Group B, each officer rating

<sup>1</sup> The authors wish to acknowledge the assistance of various members of the staff, particularly Dr. E. A. Rundquist, Mr. A. H. Birnbaum, and Mr. Joel T. Campbell. The opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Department of the Army.

10 ratees from Group A and 10 ratees from Group B. Each of the 400 officers thus received 10 such ratings, and the mean rating served as his score. The raters were informed that these ratings would be available for official use for one year, and they were required to sign the ratings. These ratings will be referred to hereafter as the signed ratings.

**Forced Choice Pairs.** The forced choice technique has been used in Army officer efficiency reports from July 1947 to September 1950. Its use in efficiency reports was discontinued largely because the form of forced choice used was not acceptable. For an extended discussion, see Baier.<sup>2</sup> Research has continued, nevertheless, with the objective of developing more acceptable forms. One of these forms contained the grouping of phrases in pairs instead of tetrads as formerly used. The phrases used were taken directly from the former Efficiency Report, Form 67-1. Twenty-four pairs were employed in this study, members of each apparently equally favorable, but differing in their ability to discriminate better from poor officers as determined by their criterion scores. The rater checked the phrase of each pair which he felt was **More Descriptive** of the ratee. Scoring was based on the discrimination values of the phrases checked. One such rating form was completed for each officer in the population by the raters of Group B.

**Controlled Check List.** The controlled check list was the second form of the forced choice technique included in this study. The phrases used in the 24 forced choice pairs were grouped into two sets of 24 phrases each. The rater selected the 12 phrases of each set which he felt were most descriptive of the ratee and scoring was based on the discrimination values of the phrases checked. One controlled check list was rendered on each officer in the population by the raters of Group A.

**Rater Variables.** Raters were divided into equal thirds, highest, middle, and lowest, on the basis of scores on each of the variables described below.

**Aptitude.** Scores on the Officer Classification Test (OCT), a high level aptitude test, were available for all officers.

**Final Class Standing.** Officer students are ranked on academic work periodically during their attendance. The last ranking is the Final Class Standing (FCS). Prior to this study, class standing and the other variables were transferred to IBM cards in such manner that the identification of individual officers was no longer possible.

**Derived Final Class Standing.** To remove the contribution of aptitude from Final Class Standing, predicted FCS was obtained from the regression of FCS on OCT. The difference between actual and predicted FCS of each officer was used as another measure of achievement referred to here as the Derived Final Class Standing (Derived FCS).

**Over-all Value.** This score was the mean of the 20 criterion rankings received by each rater.

After the raters were divided into thirds, the validity of ratings by each of these thirds on each of the rating variables was obtained. For the 8-point scales, multiple ratings per ratee were available and validities were obtained for both single ratings and mean of ratings given by each rater. For the two forced choice forms, validities were obtained for single ratings only.

## Results

The validities of ratings given by raters of each of the levels of rater characteristics studied are presented in Table 1.

In general, a direct relationship existed between the validity of the ratings given and measures of the raters' over-all value, aptitude, and achievement. Except in a few instances, validity coefficients for the highest thirds were higher than those for the middle thirds and those for the middle thirds were higher than the validity coefficients for the lowest thirds. In no case did ratings by raters in the lowest thirds have higher validities than ratings by raters in the highest thirds.

It will be noted that although the validities of ratings with the 8-point scale were slightly higher than those with the other types of ratings, the lower validity by poorer officers was

<sup>2</sup> Baier, D. E. Reply to Travers' "A critical review of the validity and rationale of the forced-choice technique." *Psychol. Bull.*, 1951, 48, 421-434.

Table 1

Validity of Ratings by Raters of Different Levels of Ability and Achievement

Raters Divided into Highest, Middle, and Lowest Thirds on		Validity of Ratings on					
		Unsigned 8-Point Rating		Signed 8-Point Rating		Forced Choice Pairs	Controlled Check List
		Single Ratings	Mean Ratings	Single Ratings	Mean Ratings	Single Ratings	Single Ratings
Aptitude (OCT)	H	.57	.74	.50	.72	.46	.45
	M	.50	.71	.51	.66	.48	.43
	L	.51	.69	.48	.60	.30	.44
Final Class Standing (FCS)	H	.58	.74	.54	.73	.50	.50
	M	.54	.71	.49	.65	.31	.46
	L	.47	.60	.44	.57	.40	.37
Derived Final Class Standing (Derived FCS)	H	.58	.74	.54	.72	.49	.55
	M	.55	.75	.50	.61	.36	.36
	L	.46	.64	.42	.62	.35	.42
Criterion Ranking	H	.61	.79	.55	.69	.49	.48
	M	.48	.66	.51	.66	.30	.48
	L	.49	.66	.45	.65	.43	.31

present in all types of scales. Of the two 8-point rating scales, the validities were somewhat higher for the unsigned than for the signed. However, Group A, which rendered the unsigned ratings, also performed the criterion ranking, so some degree of rater contamination may be present.

Both forced choice forms had greater decreases in validity from the highest third to the middle third of the rater groups than did the 8-point scales. There were three decreases as large as .19 for the forced choice forms, and the only comparably large decrease for the 8-point scales was .13. It thus appeared that raters of lesser ability, as defined in this study, produced more accurate ratings with the conventional technique than with forced choice or controlled check lists.

#### Summary

Officers at the Army Command and General Staff College rated each other using four

techniques: two 8-point scales of over-all value (one signed by the rater and the other unsigned), and two forms of the forced choice technique (forced choice pairs and a controlled check list). Rater groups were divided into highest, middle, and lowest thirds on the basis of aptitude test score, final class standing, final class standing predicted from aptitude test score, and on the criterion rank of over-all value achieved. For each third of the groups, separate validity estimates of ratings made were computed for both individual ratings and mean ratings where available.

It was found that raters who scored high on aptitude, achievement at the College, and over-all value to the Army produced more valid ratings than did raters who scored lower on these variables. This trend was highly consistent for the 8-point rating scales, and clear, though not as direct, for forced choice pairs and the controlled check list.

*Received October 15, 1952.*

# The Actuality Measure in the Study of Public Opinion

Peter R. Hofstaetter

*The Catholic University of America*

When he first suggested an "actuality-measure" (*A*) this author (7, 8, 9) aimed at combining two aspects of rather strongly ego-involved states of public opinion. He hypothesized that a population of respondents shows relatively great ego-involvement in a question if two conditions are fulfilled: (a) Few "don't know" responses; and (b) An even split between the Yes- and No-responses.<sup>1</sup>

The formula which purports to measure the actuality of a given question for a given population of respondents was based on these two notions:

$$A = \sqrt{\frac{p_+ \cdot p_-}{p_0^2}}$$

The evidence this writer has meanwhile collected seems to confirm the assumptions underlying this formula (10).

In this paper we shall first try to summarize the available evidence for the designation of *A* as a measure of "actuality" or of the involvement of a group in a topic. This is supposed to furnish an empirical check of the validity of our measure.<sup>2</sup> Our second attempt will then be to present a statistical model from which the *A*-measure can be derived and which allows us to account for the empirical properties of *A*.

After having rather regularly analyzed poll-reports during the last few years at least five conditions became clear which tend to result into relatively high actuality of questions. Each statement will be illustrated by one example; additional examples could easily be obtained.

<sup>1</sup> We shall concern ourselves only with questions that are answered in terms of a trichotomy, i.e., in the affirmative ("Yes"), in the negative ("No") or undecided ("Don't know"). The respective percentages of the responses falling in each of these categories are referred to as  $p_+$ ,  $p_-$  and  $p_0$ .

<sup>2</sup> A better check would obviously consist of an attempt to predict the relative values of *A* for questions which had been answered by different populations. Work in this direction is underway but to report on it would go beyond the scope of the present paper.

1. The more pertinent an issue is felt to be the higher is its actuality for a particular group of respondents. See Table 1.

Table 1

Distribution of Answers to the Question: "Do you think cigarette smoking is harmful or not?"†

Respondents	Harmful %	Don't know %	Not Harmful %	A
Cigarette smokers	52	3	45	16.12
Non-cigarette smokers	66	10	24	3.98

† AIPO., Dec., 1949.

2. The more imminent the event is to which the question refers the higher is the actuality of the question. See Table 2.

3. The greater the change involved in the question the higher is the actuality. See Table 3.

4. The richer the available background of experience the higher is the actuality of the question (Saenger and Gordon, 19). See Table 4.

5. The higher the educational and/or the socio-economic level of the respondents the higher actualities questions tend to attain. The data in Table 5 are based on 14 questions reported by the Psychological Corporation's Barometer (Link, 13, 14, 15, 16).

One may infer from these data that quite a few of the usual poll questions are tuned to the mentality of the above-average strata of our population. Thus they often reflect the problems of the "intelligentsia" rather than those faced by the "man on the street." There are, however, interesting exceptions to the general trend (Link, 13). See Table 6.

Cases like the one just mentioned can probably be understood in terms of point one of the present list, i.e., with reference to the greater pertinence of the question to the lower socio-economic levels.

Table 2

Distribution of Answers to the Question: "Do you think the U. S. will find itself in another world war within, say, the next year? (five years?)"†

Date of Interview	Within One Year				Within Five Years			
	Yes %	No Op. %	No %	A	Yes %	No Op. %	No %	A
May 1950	22	8	70	4.91	57	19	24	2.45
Sept. 1950	29	16	55	2.50	58	22	20	1.55
July 1951	26	10	64	4.08	56	26	18	1.22
Dec. 1952	21	12	67	3.13	48	27	25	1.28

† AIPO.

Table 3

Distribution of Answers to the Question: "Would you approve or disapprove of the following changes of postage which have been suggested?"†

Change of Postage	Per Cent Increase	Approve %	No Op. %	Disapprove %	A
Postcards from 1 ct. to 2 cts.	100	52	7	41	6.59
Reg. mail from 3 cts. to 4 cts.	33	33	8	59	5.51
Air mail from 6 cts. to 7 cts.	17	51	9	40	5.02

† AIPO., Aug., 1949.

Taken together, these examples seem to suggest the empirical validity of the actuality-measure. They represent, however, only a small sample from the total available evidence.

While the empirical validity of the actuality-measure seems to stand on relatively firm ground the arbitrariness of its derivation caused the present writer considerable headache. It will be the task of the present discussion to remedy this shortcoming.

Let us assume, for the sake of a model, that our respondents were to give their answers on the basis of two balls drawn from

two urns. Each of these urns contains "a" per cent white and "b" per cent black balls. The following combinations can thus occur: two white balls, one white and one black ball, two black balls. According to the rules of our game the respondents will have to render a "Yes-statement" whenever they draw two white balls, and a "No-statement" if both balls are black. In case they draw one white and one black ball the answer is going to be "Undecided," or "Don't know," "Uncertain," "No opinion," etc.

The frequencies with which we expect these

Table 4

Distribution of Answers to the Question: "Do you believe in the efficiency of the New York State law against discrimination?"

Respondents	Yes %	Don't Know %	No %	A
Persons who had experienced job discrimination	15	14	71	2.33
Others	27	21	52	1.78

Table 5

Actuality as a Function of the Socio-economic Level of the Respondents

Socio-economic Level	Average A	Average A (%) (Total = 100)
A (highest 10 per cent)	4.91	151
B (next 30 per cent)	4.04	131
C (next 40 per cent)	3.28	100
D (lowest 20 per cent)	2.03	61
Total	3.32	100

Table 6

Distribution of Answers to the Question: "Have businessmen a right to shut down?"

Socio-economic Level	Yes %	Un-certain %	No %	A	A %
A	60	8	32	5.48	83
B	57	7	36	6.48	98
C	46	7	47	6.64	100
D	41	6	53	7.78	117
Total	49.5	8	43.5	6.63	100

responses to occur will follow the binomial expansion:

$$(a + b)^2 = a^2 (\text{Yes}) + 2ab (\text{Don't know}) + b^2 (\text{No}).$$

The middle term in this expansion becomes thus:

$$2ab = 2\sqrt{a^2b^2} = p_0.$$

We replace now  $a^2$  and  $b^2$  by the observed percentages of the Yes- and No-responses ( $p_+$  and  $p_-$  respectively) and thus obtain for the middle term:

$$p_0 = 2\sqrt{p_+p_-}.$$

For all binomial distributions the following relationship holds:

$$P = \frac{2\sqrt{p_+p_-}}{p_0} = 1.00.$$

This provides us with an easy way to ascertain whether or not an empirical distribution follows the binomial pattern. We have to reject the simple binomial (or "Bernoullian") model whenever the ratio  $P$  differs significantly from 1.00.

The magnitude  $P$  becomes thus a measure for the applicability of the Bernoullian model to an observed distribution of responses. It can, however, be shown that our actuality-measure ( $A$ ) serves this very same purpose equally well:

$$P^2 = \frac{4p_+p_-}{p_0^2} = 4A^2; \text{ consequently: } A = \frac{1}{2}P.$$

Since  $P$  equals 1.00 for all binomial distributions  $A$  necessarily equals 0.50.

It thus becomes clear that our actuality-

Table 7

Distribution of Answers to the Question: "Do you think that Socialism in England will succeed or fail?"

Socio-economic Level	Succeed %	Don't Know %	Fail %	A	A %
A	19	24	57	1.37	236
B	13	31	56	0.87	150
C	10	42	48	0.52	89
D	11	60	29	0.30	52
Total	12	41	47	0.58	100

measure is nothing but a quantitative expression for the appropriateness of the Bernoullian model for the representation of an empirical distribution of responses. It should be noted that this model does not presuppose any specific proportionality between the two initial probabilities "a" and "b." The appearance of a rather substantial majority for either "Yes" or "No" is therefore compatible with the Bernoullian model.

The examples given in the first part of this paper have shown rather consistently actualities higher than 0.50. Indeed, it is not easy to locate questions which yield actualities that low or lower. An example may be drawn, however, from one of Link's polls (14). See Table 7.

The response distribution of Group C ( $A = 0.52$ ) comes indeed very close to the expansion of  $(0.30 + 0.70)^2 = 9\%$  (Succeed) +  $42\%$  (Don't know) +  $49\%$  (Fail).

Since poll data conform only rarely to this pattern we have to recognize that the Bernoullian model is too simple to account for them. But we know already that its failure in that respect can be gauged from either  $P$  or  $A$ .

The next step we have to undertake is to replace the model of two urns with the same probabilities ("a" and "b") in each by the model of two urns with different probabilities. Obviously, our first model can be considered as a special case of this second, more general model.

Let us assume that the respondents base their decisions on the drawing of one ball from each of two urns where the following

Table 8

The Markov Model of a Response Distribution

Urn	Probabilities			A
I	$a_1=0.50$	$b_1=0.50$		
II	$a_2=0.70$	$b_2=0.30$		
III	$a_3=0.10$	$b_3=0.90$		
Responses				
%	$p_+=35$	$p_0=20$	$p_-=45$	1.98

probabilities prevail:

$$a_1 \neq a_2 \text{ and } b_1 \neq b_2.$$

The corresponding percentages of the three categories of answers are:

$$\begin{aligned} p_+ &= a_1 \cdot a_2 \\ p_0 &= (a_1 \cdot b_2) + (a_2 \cdot b_1) \\ p_- &= b_1 \cdot b_2 \end{aligned}$$

This model always results in distributions which have actualities of less than 0.50. Its usefulness is therefore drastically limited.

This will no longer be the case with the third model which employs the statistical theory of the "Markov Chains" (5). Let us assume that there are three urns with different probabilities for white and black balls:

$$a_2 > a_1 > a_3 \text{ and } b_2 < b_1 < b_3.$$

Whenever a subject has drawn a white ball from Urn I he is bound to draw his second ball from Urn II where the probability of a white ball is even greater than in the first urn. In the event of a black ball being drawn from the first urn the subject is required to draw the second ball from Urn III which gives him again an even higher probability to draw a black ball than the first urn did. This model abandons the independence condition of the Bernoullian model. A numerical example may help to clarify this notion. See Table 8.

The essential feature of this model is that the probabilities which prevail for the drawing of the second ball depend on the result of the first drawing. Such probabilities have been called "dependent" or "conditional" probabilities. It can be seen easily that the Markov model includes the Bernoullian model as a special case of  $a_1 = a_2 = a_3$ . The out-

come of a so-called Markov process is defined by the following equations:

$$\begin{aligned} p_+ &= a_1 \cdot a_2 \\ p_0 &= (a_1 \cdot b_2) + (b_1 \cdot a_3) \\ p_- &= b_1 \cdot b_3 \end{aligned}$$

Though it may seem as if we had three equations for the determination of the three unknown variables ( $a_1$ ,  $a_2$  and  $a_3$ ) we will not be able to solve these equations since only two of the  $p$ -values are independent. The problem is therefore bound to remain underdetermined unless we introduce an arbitrary assumption with respect to the initial probability  $a_1$ . Our assumption may, for instance, be:  $a_1 = b_1 = 0.50$ .<sup>3</sup> Consequently, the following equations will define the conditional probabilities:

$$\begin{aligned} a_2 &= \frac{1}{a_1} p_+ = 2p_+; \\ a_3 &= \frac{p_0 - a_1 b_2}{b_1} = 2(p_+ + p_0) - 1.00. \end{aligned}$$

Since the observational data we collect by means of poll questions do not allow us to determine all three probabilities needed for the Markov model its usefulness may seem to be sort of academic. Yet, it should be pointed out that this model shows one property which furthers our understanding of the processes which underlie the formation of public opinion. In order to become aware of this aspect we have only to replace the term "urn" by the term "source of information." The Markov model tells us that the conclusions our respondents draw from one source of information (these conclusions may be either favorable or unfavorable with respect to the question at issue) determine their exposure to other sources of information (some of them being more and others less likely to confirm a favorable attitude). What accounts ultimately for an observed response distribution is thus a certain amount of selectivity with respect to the sources of information to which the respondents expose themselves. What this boils down to is that the drawing from different sources of information does not occur in a statistically independent manner. On

<sup>3</sup> This assumption becomes untenable in the case of:  $p_+ + p_0 < 0.50$ .

the contrary, the sources of information which our respondents actually recognize are in most cases intercorrelated.

Once the first choice has occurred the subsequent choices tend to be made on the basis of sources of information which are more likely to support the initial choice than to contradict it. Thus a series of choices gets underway which assures the individual subject of a relatively stable and un-ambivalent attitude. By the same token we may infer that neutral responses ("Don't know," etc.) tend to vanish in the group of respondents as a whole. What we have called "actuality" and have operationally defined by our formula thus serves as an indicator for the amount of "bundling" of sources of information which prevails in a given group of respondents with respect to a given issue.

Table 9

The Markov Model of Two Correlated Sources

Before Debate				After Debate			
$p_+$	$p_0$	$p_-$	$A$	$p_+$	$p_0$	$p_-$	$A$
36	46	18	0.55	36	18	46	2.26
Urn II				Urn II			
Urn I	$a_2$	$b_2$	Sum	$a_2$	$b_2$	Sum	
$a_1$	36	23	59	36	9	45	
$b_1$	23	18	41	9	46	55	
Sum	59	41	100	45	55	100	
$r_t = 0.08$				$r_t = 0.83$			

In order to explore this last point further we can restate the Markov model in the following way: We assume two sources of information ("urns") with the same probabilities for either positive or negative statements in each of them:  $a_1 = a_2$ ;  $b_1 = b_2$ . We can thus set up a fourfold table in which all sorts of trichotomic response distributions can be arranged as is shown in Table 9.<sup>4</sup>

The data of Table 9 are taken from an experiment by Millson (17) in which the author arranged for group discussions on the topic of unemployment insurance. Before and again after debate the participants were

<sup>4</sup> The tetrachoric correlation coefficients ( $r_t$ ) are read from the well-known Thurstone Tables (3).

asked to state their opinions. As can be seen from Table 9 the debate resulted in a heightened actuality of the question. This is a fairly common event (10). In the lower half of Table 9 the same data have been brought in the form of fourfold tables in order to allow for the computation of tetrachoric correlations. The value of  $p_0$  had thus to be divided into two cells, i.e., the combination of  $a_1$  with  $b_2$  and of  $a_2$  with  $b_1$ . One half of  $p_0$  was allotted to either cell.

The interpretation that can be given to these data reads: Whereas the respondents used the sources of information available to them before the debate in an (almost) independent manner a considerable degree of bundling or patterning occurred after the debate. The rise in actuality (from 0.55 to 2.26) corresponds with an increase in the degree of bundling (from  $r_t = 0.08$  to  $r_t = 0.83$ ). In fact, there exists, in general, a definite relationship between  $A$  and the corresponding coefficient of correlation  $r_t$ . See Figure 1.<sup>5</sup>

Actualities below 0.50 correspond to negative correlations. This means that the respondents who have drawn their conclusions from the first source of information in a positive manner are more likely to draw a negative conclusion from the second source they consult, and vice versa. This is seldom the case. An example can be seen in the behavior of socio-economic group D in Table 7. In introspective terms this may be either considered as "disappointment" or as a striving for objectivity.

As we reach the end of our present investi-

<sup>5</sup> An empirical equation has been fitted to this curve:

$$r_t = 0.96 - 1.78 e^{-1.093 A}$$

For  $A$ -values over 0.50 the agreement between expected and observed  $r_t$ -values is excellent, below 0.50 it is tolerable. It had been hoped that this relationship would provide us with a test of significance for  $A$ . This hope may, however, not materialize because of the great difficulties that enter into the determination of the significance of differences between tetrachoric coefficients. In the future a significance test for  $A$  in terms of the corresponding Chi-squares seems to be more likely. It can, however, be reasonably inferred from Figure 1 that relatively small numerical differences between  $A$ -values in the range from 0.50 through 3.50 are much more likely to be significant than numerically larger differences between higher  $A$ -values.

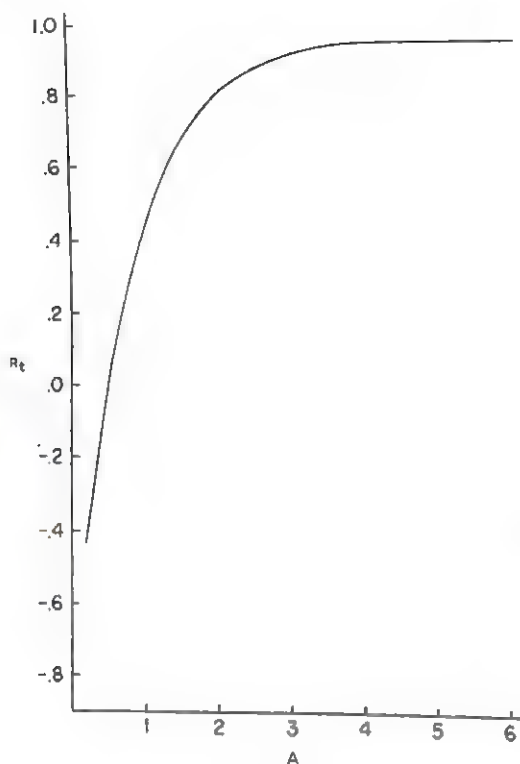


FIG. 1. The relationship between actuality ( $A$ ) and the bundling of sources of information ( $r_t$ ).

gation it may be said that the actuality-measure indicates the amount of discrepancy that exists between an observed response distribution and the kind of distribution that one would expect on the basis of the Bernoullian model. This discrepancy can be expressed in terms of the correlated functioning of two sources of information (Markov model).

Needless to say, no model can ever be taken literally. It is quite artificial to argue that always two and just two sources enter into the determination of a person's attitude towards a given social issue. Indeed, there is little doubt that we are exposed to many more sources of information with respect to most issues.<sup>6</sup> It seems, however, that the more sources of information the ordinary person, in accordance with his socio-economic and/or educational level, consults, the

stronger becomes his chance to hit upon correlated sources. This is in line with point five in the first part of this paper. The other four points previously made can now be summarized by saying that the more important a decision we have to make, the more prone we are to expose ourselves to correlated sources of information. Yet, this statement requires at least one qualification. As has been shown in an earlier paper (9) there exists a negative correlation between the actuality and the experienced difficulty of questions. Our subjective experience that a question is difficult seems to correspond with an un-bundling of sources of information. Ultimately this causes us to shift in the direction of the Bernoullian model.

To recognize fully the eventually large number of sources of information would, of course, complicate our derivations. The basic postulate, however, will probably remain unaltered, that is, that one can either use the multitude of sources of information in a complete and unselected manner or use them selectively. From a theoretical point of view the first procedure is undoubtedly preferable; in practice the latter seems to be the rule. To follow this "rule" minimizes the ambiguity of perceived situations and thus eases the burden of uncertainty.<sup>7</sup> *Ceteris paribus* this tendency may be expected to show up more strongly the more deeply ego-involved we are in certain issues. Actuality measures the strength of this tendency.

### Summary

1. Poll questions which can be answered by either "Yes," "No" or "Don't know" yield response distributions which can be compared

<sup>7</sup> This very same mechanism has been described in several situations by different terms. For instance as the "selectivity" of learning and forgetting (5, 12), or as "intolerance of ambiguity" (6) with regard to perception. The relationship between "intolerance of ambiguity" and ethnocentrism has been demonstrated (1). A comparable tendency to escape from ambiguity seems to underlie the "Halo-Effect" (Thorndike) in personality judgment. The correlated use of indicators refers also to this situation (2). H. Werner's (20) concept of "rigidity" is also based upon the notion of either too much isolation or too much overlapping of subareas, i.e., upon their positive or negative correlation. The "hypothesis-confirmation theory" of Postman (18) provides a suitable frame of reference for these phenomena.

<sup>6</sup> The "interference-hypothesis" which asserts "that any communication which successfully modifies a person's beliefs will reduce the opinion-impact of any subsequent event or communication that tends to produce antithetical beliefs" has been recently confirmed (11).

with binomial expansions.

2. The actuality-measure indicates the discrepancy between an observed response distribution and that expected from the binomial model.

3. All kinds of trichotomic response distributions can be fitted to the Markov model of dependent probabilities.

4. Under simplifying conditions, the actuality-measure becomes an indicator of the amount of bundling (or correlation) among the sources of information to which the respondents have exposed themselves.

5. The bundling of sources of information increases with the subjectively felt importance of the issue to which the question relates and also—in general—with the socio-economic and/or the educational level of the respondents. It decreases with the subjectively felt difficulty of the issue.

Received August 20, 1952.

#### References

1. Block, J. and Block, Jeanne. An investigation of the relationship between intolerance of ambiguity and ethnocentrism. *J. Pers.*, 1951, 19, 303-311.
2. Brunswik, E. *Systematic and representative design of experiments*. Berkeley: Univ. of Calif. Press, 1947.
3. Chesire, L., Saffir, M., and Thurstone, L. L. *Computing diagrams for the tetrachoric correlation coefficient*. Chicago: Univ. of Chicago Bookstore, 1933.
4. Edwards, A. L. Political frames of reference as a factor influencing recognition. *J. abnorm. soc. Psychol.*, 1941, 36, 34-61.
5. Feller, W. *An introduction to probability theory and its applications*. New York: Wiley, 1950.
6. Frenkel-Brunswik, E. Intolerance of ambiguity as an emotional and perceptual personality variable. *J. Pers.*, 1949, 18, 108-143.
7. Hofstaetter, P. R. *Die Psychologie der öffentlichen Meinung*. Wien: Braumueller, 1949.
8. Hofstaetter, P. R. The actuality of questions. *Int. J. Opin. Attitude Res.*, 1950, 4, 16-26.
9. Hofstaetter, P. R. Importance and actuality. *Int. J. Opin. Attitude Res.*, 1951, 5, 31-52.
10. Hofstaetter, P. R. *Einführung in die Sozialpsychologie*. Wien: Humboldt, 1953.
11. Janis, I. L., Lumsdaine, A. A., and Gladstone, A. I. Effects of preparatory communications on reactions to a subsequent news event. *Publ. Opin. Quart.*, 1951, 15, 487-518.
12. Levine, J. M., and Murphy, G. The learning and forgetting of controversial material. *J. abnorm. soc. Psychol.*, 1943, 38, 507-517.
13. Link, H. C. The Psychological Corporation's index of public opinion. *J. appl. Psychol.*, 1946, 30, 297-309.
14. Link, H. C. The ninety-fourth issue of the Psychological Barometer and a note on its fifteenth anniversary. *J. appl. Psychol.*, 1948, 32, 105-117.
15. Link, H. C., and Freiberg, A. D. The ninety-seventh Psychological Barometer. *J. appl. Psychol.*, 1948, 32, 443-451.
16. Link, H. C., and Freiberg, A. D. The Psychological Barometer on Communism, Americanism and Socialism. *J. appl. Psychol.*, 1949, 33, 6-14.
17. Millson, W. A. D. Problems in measuring audience reaction. *Quart. J. Speech*, 1932, 18, 621-637.
18. Postman, L. Toward a general theory of cognition. In J. H. Rohrer and M. Sherif (Eds.), *Social psychology at the crossroad*. New York: Harper, 1951.
19. Saenger, G., and Gordon, N. S. The influence of discrimination on minority group members in its relation to attempts to combat discrimination. *J. soc. Psychol.*, 1950, 31, 95-120.
20. Werner, H. The concept of rigidity; a critical evaluation. *Psychol. Rev.*, 1946, 43, 43-52.

## A Biasing Factor in Essay Response Frequency

Erwin K. Taylor and Dorothy E. Schneider

*Personnel Research Institute, Western Reserve University*

In order to get descriptive phrases for use in developing an evaluation form, questionnaires were mailed to 955 randomly selected members of a professional organization, the American Dietetic Association. The questionnaire explained the purpose of the study and asked each respondent to describe an associate at a level of competence specified on the form. Each questionnaire specified a single level out of ten possible levels. Ninety-five or ninety-six questionnaires were sent out for each level. Of the ten levels, "1" was considered best, and "10" poorest. Although there was some feeling on the part of the present writers that the returns, being on an anonymous and voluntary basis, would yield either a J- or U-shaped distribution, there was no published evidence to justify unequal distribution in mailing of the forms. Hence, each level was requested from an equal number of recipients of the questionnaire.

In addition to the descriptive essays, respondents were asked a number of questions regarding the type of work engaged in when they were acquainted with the dietitian de-

scribed, length of acquaintance, and working relationship between the two.

The questionnaires were mailed from Cleveland on June 19, 1952; the first reply was received on June 24, and by July 10, the "deadline" requested on the questionnaire form, 130 forms had been returned. Within the two weeks following, 14 more were received, bringing the total to 144, or about a 15% return.

The forms which were returned reveal some interesting points. The distribution at each level of competence is shown in Table 1, according to whether the respondent selected a subordinate, co-worker, superior, or student as subject for the essay.

It will first be noted that the distribution, while not horizontal, does not reveal sufficient U or J shape to be so named. The mean of 5.4 is very close to the 5.5 which would be expected in a horizontal distribution. At only two levels, 4 and 6, are the frequencies more than 5 below the maximum frequency.

A Chi-square test was applied to the total frequencies and revealed no statistically significant deviation from the theoretical recti-

Table 1  
Distribution of Returns on Questionnaire Form Requesting Descriptive Essays

Level of Competence of Person Described	Relation of Subject to Writer of Essay				Total
	Sub- ordinate	Co- Worker	Superior	Student	
1 (Best)	6	0	10	2	18
2	5	3	10	0	18
3	5	4	4	1	14
4	1	2	5	2	10
5	4	5	7	0	16
6	4	3	2	0	9
7	3	6	3	1	13
8	9	2	2	1	14
9	4	4	4	2	14
10 (Poorest)	9	5	4	0	18
Total	50	34	51	9	144
Mean	5.96	6.26	4.35	5.11	5.40

linear distribution. This indicates the absence of bias in the returns in terms of the level of competence of the individual to be described.

Thus, while it has been conjectured that it is easier to evaluate personnel at the extremes of the effectiveness distribution, it does not appear to be necessary to take this factor into account in planning the collection of data of this nature. Responses were received with almost equal frequency for ratees at each of the ten effectiveness levels used in this study.

When no restriction was placed upon the selection of subjects for the essays, other than the level of competence, about the same number of supervisors and subordinates were selected. However, some rather striking differences will be observed in the mean levels of competence of the two groups. It would seem

that the more competent "superiors" were remembered by the recipients of the questionnaire and selected for description, while those who were to describe less competent individuals tended to select subordinates.

It is interesting to note that no respondent who was asked to describe a No. 1 person—the best—described a co-worker, and that fewer co-workers were described than either superiors or subordinates. Although there is the same proportion of co-workers as subordinates in levels 6–10, the mean is lower. This is largely a function of the zero frequency at level 1.

The number of descriptions of students is small, but no unusual trends are noted in manner of selection.

*Received October 15, 1952.*

# Rating Patterns for Maximizing Competition and Minimizing Number of Comparative Judgments Necessary for Each Rater

Ray H. Simpson

*Department of Educational Psychology, University of Illinois*

The problem considered in this paper is represented by a situation which contains the following features:

a. A large number of individuals or written products are to be ranked.

b. The time required to evaluate and give a rank to each individual or product is of such magnitude that any single judge cannot be asked to rank more than 3, 4, or 5 individuals or products.

c. The researcher or administrator wants to use a large number of judges in order to maximize the reliability of the combined rating or ranking.

d. Each individual or product should directly compete with as many other individuals or products as possible.

e. The total group of individuals or products with which a particular individual or product competes should represent a random sample of the total group of competitors against whom he is to be ranked.

The method of paired comparisons would obviously not be appropriate since it would take too much time and be too fatiguing to the judges. Forty individuals or products to be ranked, for example, would involve  $n(n-1)/2$  or 780 pairs to be compared by each judge.

Consideration was given to the possibility of using a technique suggested by Guilford<sup>1</sup> which involves selecting from all the individuals or products a limited number to become the basis for a scale. This alternative was not used because of the difficulty in selecting appropriate products or individuals at approximately equal intervals along the scale and because of the volume of work still required of each judge.

The technique described by Uhrbrock and Richardson<sup>2</sup> was also considered. In it one

breaks the individuals to be rated into four groups. Within each group each man is compared with every other man in his group and also with each of five "key" men. These latter five form a type of interlocking yardstick. While this method is much more economical of the time of judges than the paired comparison technique, it still was too time consuming to meet all of the requirements of the writer's situation. As a result, the patterns described below have been developed.

## Rating Patterns

Table 1 illustrates a pattern to be used when: (a) there are 21 individuals or products to be ranked from highest to lowest; (b) there are to be 21 raters (A through U); (c) each individual or product is to compete with

Table 1

Interlocking Design for Having 21 Raters Give Competitive Ranks to 21 Individuals or Products

Rater	Individuals or Products to be Rated				
A	1	2	7	9	19
B	2	3	8	10	20
C	3	4	9	11	21
D	4	5	10	12	1
E	5	6	11	13	2
F	6	7	12	14	3
G	7	8	13	15	4
H	8	9	14	16	5
I	9	10	15	17	6
J	10	11	16	18	7
K	11	12	17	19	8
L	12	13	18	20	9
M	13	14	19	21	10
N	14	15	20	1	11
O	15	16	21	2	12
P	16	17	1	3	13
Q	17	18	2	4	14
R	18	19	3	5	15
S	19	20	4	6	16
T	20	21	5	7	17
U	21	1	6	8	18

<sup>1</sup> Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.

<sup>2</sup> Uhrbrock, R. S. and Richardson, M. W. Item analysis. *Personnel J.*, 1933, 12, 141-154.

Table 2

Abbreviated Rating Patterns with Each Judge Rating Five Individuals or Products

Total Number to be Rated—25					
Judge	Rates				
A	1	2	4	8	16
B	2	3	5	9	17
C	3	4	6	10	18
.	.	.	.	.	.
.	.	.	.	.	.
W	23	24	1	5	13
X	24	25	2	6	14
Y	25	1	3	7	15
Total Number to be Rated—26					
Judge	Rates				
A	1	2	4	8	13
B	2	3	5	9	14
C	3	4	6	10	15
.	.	.	.	.	.
.	.	.	.	.	.
X	24	25	1	5	10
Y	25	26	2	6	11
Z	26	1	3	7	12
Total Number to be Rated—27					
Judge	Rates				
A	1	2	4	8	17
B	2	3	5	9	18
C	3	4	6	10	19
.	.	.	.	.	.
.	.	.	.	.	.
Y	25	26	1	5	14
Z	26	27	2	6	15
AA	27	1	3	7	16
Total Number to be Rated—28					
Judge	Rates				
A	1	2	4	8	13
B	2	3	5	9	14
C	3	4	6	10	15
.	.	.	.	.	.
.	.	.	.	.	.
Z	26	27	1	5	10
AA	27	28	2	6	11
BB	28	1	3	7	12
Total Number to be Rated—29					
Judge	Rates				
A	1	2	4	8	18
B	2	3	5	9	19
C	3	4	6	10	20
.	.	.	.	.	.
.	.	.	.	.	.
AA	27	28	1	5	15
BB	28	29	2	6	16
CC	29	1	3	7	17
Total Number to be Rated—30					
Judge	Rates				
A	1	2	4	8	22
B	2	3	5	9	23
C	3	4	6	10	24
.	.	.	.	.	.
.	.	.	.	.	.
BB	28	29	1	5	19
CC	29	30	2	6	20
DD	30	1	3	7	21
Total Number to be Rated—31					
Judge	Rates				
A	1	2	4	8	16
B	2	3	5	9	17
C	3	4	6	10	18
.	.	.	.	.	.
.	.	.	.	.	.
CC	29	30	1	5	13
DD	30	31	2	6	14
EE	31	1	3	7	15
Total Number to be Rated—32 <sup>1</sup>					
Judge	Rates				
A	1	2	4	8	16
B	2	3	5	9	17
C	3	4	6	10	18
.	.	.	.	.	.
.	.	.	.	.	.
DD	30	31	1	5	13
EE	31	32	2	6	14
FF	32	1	3	7	15

<sup>1</sup> Similar designs with the same initial sequence can be used with groups larger than 32.

as many other individuals or products as possible; and (d) no judge or rater is to be asked to rank more than five individuals or products. The time involved in the job of rating, reading, or observing necessitates this last requirement.

Examination of Table 1 will show that individual or product number 9, for example, is in competition with each of the 20 other numbers. This means that the final rank for number 9 is as fair a rating as one can get without excessive work on the part of individual judges.

This particular design was used by the writer in a class of 21 students where each student studied and rated, on the basis of

criteria previously established, the learning products of five peers. For each student to rank more would have been too fatiguing. The sum of all ranks assigned a particular individual was used to rank the 21 students in the class. Other patterns, based on similar designs, are represented in essence in Table 2.

The technique and patterns suggested here have wide applicability. A few of the additional areas where use would be feasible would be the following: ranking personnel such as in an industrial, military or educational situation; rating short stories; and ranking art forms for esthetic qualities.

*Received September 2, 1952.*

## Some Primary Ratable Characteristics of Instructional Films<sup>1</sup>

Philip Ash<sup>2</sup> and Thelma R. Hobaugh

*The Pennsylvania State College*

A central problem facing the instructional film producer on the one hand, and the teacher on the other, is the evaluation of the film as to its adequacy as a teaching aid.

The *primary* objective of an instructional film is, generally, to instruct. Therefore, the first criterion of the adequacy of a film as an instructional device must be: how well does the film communicate its content—how much is learned from it?

However, a second important objective of an instructional film is to interest, to hold attention, possibly to entertain in the broad sense. In some of its uses, for example as an overview to a unit of instruction, the teacher might intend that the film motivate the students *to want to learn* about the content, with little emphasis on information directly acquired from the film. Furthermore, it seems likely that the extent to which a film is "interesting" in the sense that it gains and holds the attention of the students will influence the extent to which it communicates effectively.

Three kinds of estimates of the quality of an instructional film therefore suggest themselves: first, measures of learning based on test results; second, estimates (judgments, ratings) by people who have seen the film as to the extent to which people will learn from the film; and third, ratings of the affective impact of the film more or less independently

of its effectiveness in imparting factual information.

It is the purpose of the present research to determine the extent to which these kinds of estimates are intercorrelated, to answer the question as to the extent to which, from a *rating* of teaching effectiveness or affective quality, it is possible to predict *measured learning*.

A rough check on the teaching effectiveness of an instructional film may be made by administering, both before and after the film, a test generally covering the content of the film. Gains in test scores would reflect the over-all contribution of the film to the students' fund of knowledge. Such a test, however, would not clearly indicate which *parts* of the film were effective and which were ineffective.

A better method of evaluation might involve the development of a *profile measurement* which shows which *parts* of the film are effectively imparting information and which are less effective or ineffective in this respect. Furthermore, techniques are available for obtaining rating profiles of the "attention value" and other affective attributes of a film, and for obtaining judgments of teaching value. Profile methods of summing audience ratings with respect to successive sections of programs (entertainment films, radio shows) have been developed extensively during the past ten years, primarily in the radio and motion picture industries.<sup>3</sup> Little comparable work has been done with instructional films, however.

### Film Profile Analysis

Under the auspices of the Instructional Film Research Program at The Pennsylvania State College, Dr. Loran C. Twyford conducted an extensive study of the feasibility of a variety of methods for developing profiles of audience evaluations of the instructional

<sup>1</sup> The research on which this article is based was conducted under the auspices of the Instructional Film Research Program under Contract N6-ONR-269, Task Order VII with the Special Devices Center of the Office of Naval Research. This research is reported in Technical Report SDC 269-7-23, *Film Profiles*. The authors gratefully acknowledge permission to analyze the correlation matrix that evolved out of the Ph.D. dissertation completed by Loran C. Twyford: *A Comparison of Methods for Measuring Profiles of Learning from Instructional Films*, The Pennsylvania State College, 1951; Publication No. 3314, University Microfilms, \$2.76. Ann Arbor, Michigan. Dr. Ash was responsible for planning the study and writing the report; Miss Hobaugh was responsible for carrying out the factor analysis.

<sup>2</sup> Dr. Ash is now a member of the Industrial Relations Research Department, Inland Steel Company, East Chicago, Indiana.

<sup>3</sup> An extensive review of profile analysis techniques in radio and motion pictures is included in Dr. Twyford's study, *op. cit.*

film. These profiles were graphs, drawn against time as the base-line, of the summed responses of the audience to several dimensions of the film.

The profiles he collected may be classified into three groups:

(1) *Profiles of measured learning and learning gains.* These profiles were based upon an exhaustive test of the film content. The test items were true-false form; every fact presented in the film was covered by an item, and each item could be related back to a specific point (in time, sequence, and footage) in the film.

(2) *Profiles of estimated teaching effectiveness.* For these profiles the subjects were asked: (a) to rate whether they thought they were learning; (b) to rate whether they thought students would learn; or (c) to rate (on a third showing) whether they had learned the content presented. The ratings were made during the film showing, on a five-point scale. The subjects made their responses by means of the Film Analyzer<sup>4</sup> system, which provides for the collection by polygraphic recording of the responses of each of a group of 40 subjects.

(3) *Profiles of affect dimensions.* Profiles were collected on the dimensions: (a) "Like-Dislike the film"; (b) "The film is Clear-Unclear"; and (c) "The film is Good-Bad." These profiles also were made on five-point scales and the judgments were recorded by means of the Film Analyzer system concurrently with the viewing of the film.

### Procedures

In all, Dr. Twyford collected data on 276 high-school and college students, divided into seven equated groups.

The students in the experiment met as intact classes. A given class was then randomized into seven groups. Each of the seven groups in a

<sup>4</sup> At each of the 40 stations (chairs) of the Film Analyzer system a box containing five piano-like keys is located. Each subject presses a key signifying his response, and a simultaneous record of the 40-response channels is made on a moving paper. The five choices are coded so that each response channel produces dashed, solid, or a pair of dashed and solid lines identifying the key being pressed. For a description, see Carpenter, C. R., et al. *The Film Analyzer*. Special Devices Center Technical Report 269-7-15, Instructional Film Research Program, The Pennsylvania State College, 1950.

class followed a different one of seven experimental procedures.

The film shown to the experimental population was a 10-minute section of a longer film on precision measuring instruments. The original film had been designed by the Instructional Film Research Program. It is essentially an informational film which undertakes to show various kinds of precision measuring instruments and their method of use. The 10-minute section used in the experiment was a complete and intact section of the original.

Groups I, II, and III each saw the film three times, and made a continuous rating on each occasion. Group IV took Form A of the test after seeing the film once, Group V took Form B after seeing the film once. Groups VI and VII took Forms A and B respectively without seeing the films.

The experimental directions were (in part) as follows:

1. "*I am learning.*" Specific instructions given to students making this rating were:

"You are to rate the film on the basis of your learning from the film. When facts are presented by the film that you already know, you should report that you are not learning. Some facts in the film may be so difficult that you could not answer questions about them if you were asked to do so after the film showing. These facts were not learned. Learning means that some change has occurred to you that we could discover by comparing your knowledge before and after the film showing."

2. "*I predict learning.*" Specific instructions given to students making this rating were:

"You are to predict the amount of learning you would expect a group composed of your classmates to learn from a single showing of the film. The group will not be learning much if the material being presented is previously known. The group will not be learning if the material being presented is too difficult. Learning may occur when the film presents facts that are new. Learning means that some change has occurred to the person that we could discover by comparing his knowledge before and after the film showing by means of tests."

3. "*I like-dislike.*" Students making this rating were given the following instructions:

"You are to rate the film according to the degree to which you like the parts of the film. If you like a part very well, depress key No. 5. If you dislike a part of the film, depress key No. 1. The keys that lie between should be pressed to show just how much you like or dislike each part. It will be possible for us to improve the film by making the entire film similar to the parts that you indicate you like. Rate the film from beginning to end."

During a second showing of the film to the same groups, each group was asked to respond a

second time, using the same judgment that had been used the first time.

During the third showing, the students in one group were asked to rate how much they had learned; in the second group they were asked to rate how clear the film was; in the third group they were asked to rate the excellence of the film as a teaching aid.

The remaining four groups of each class were handled as follows:

One group took form A of a true-false test which exhaustively sampled the information contained in the film (in both pictures and commentary).

A second group took form B of the same test (true questions of form A were false questions in form B; false questions in form A were true questions in form B).

These two groups acted as a control on the pre-film knowledge of the experimental population; they did not see the film.

The two remaining groups received no pre-tests. They were each shown the film, and then were respectively given one of the two test forms mentioned above.

The nine rating profiles (Groups I-III) were each established by taking the average rating for the group for every two feet of film, a total of 151 points for each profile.

The test questions were distributed approximately evenly to provide a measure in each interval. The four basic test profiles (for Groups IV-VII) were prepared by calculating the mean score for each point on the time axis. Some questions were used at more than one point in the film. These four basic profiles generated a variety of profiles of learning gains (difference between a post-test profile and an appropriate pre-test profile).

The list of the profiles is as follows:

1. Footage—number of feet of elapsed film, measure of the time base.
2. I am learning—1st showing: Ratings of extent of learning during first viewing of film by Group I.
3. Like-Dislike—1st showing: Ratings of degree to which viewer liked the film, during first viewing of film by Group II.
4. Predicted learning—1st showing: Ratings of the extent to which one's classmates would learn the content, during the first viewing of the film by Group III.
5. I am learning—2nd showing: As for profile 2, during second viewing by Group I.
6. Like-Dislike—2nd showing: As for profile 3, during second viewing by Group II.
7. Predicted learning—2nd showing: As for profile 4, during second viewing by Group III.
8. Rated end knowledge—3rd showing: Ratings of amount learned during previous showings, during third viewing by Group I.

9. Clarity: Ratings of the extent to which the film presentation was clear and understandable, during third viewing by Group II.
10. Good-Bad film: Ratings of the excellence of the film as a teaching aid, during third viewing by Group III.
11. Test A learning: Mean score of Group IV minus mean score of Group VI (control or no-film group for Form A) for each profile point.
12. Test B learning: Mean score of Group V minus mean score of Group VII (control or no-film group for Form B) for each profile point.
13. True questions learning: Mean score on true questions from Groups IV and V minus mean score on true questions from Groups VI and VII.
14. False questions learning: As for profile 13, but for false questions.
15. Comparable questions learning I: Mean score for Group V minus mean score for Group VI, for comparable questions.
16. Comparable questions learning II: As for profile 15, Group IV minus Group VII.
17. Maximum learning: Profile obtained by assuming maximum learning.
18. Terminal knowledge, Form A: Group IV mean scores for profile points.
19. Terminal knowledge, Form B: As for profile 18, for Group V scores.
20. Terminal knowledge, true questions: Group IV mean scores and Group V mean scores on true questions only.
21. Terminal knowledge, false questions: As for profile 20, for false questions.

### Purpose of Present Analysis

Dr. Twyford was primarily interested in developing profile construction methodology, and in comparing the relative efficiency, with *measured learning* as a criterion, of the *affected rating* profiles and the *estimated teaching effectiveness* profiles for detecting the strong and weak points of the film. Part of his analysis, therefore, involved correlating the profiles. This was done by taking, for each profile, a measurement at every two feet of film length (151 measurements in all) and intercorrelating the profile arrays.<sup>5</sup>

<sup>5</sup> Before further analysis, Dr. Twyford applied a series of corrections to his correlations. These corrections were not employed in this analysis. These corrections for "drift," "lag," and "carry-over" involved either smoothing lines or partialling out time. They were not applied in this factor analysis; first, because they had little effect on the observed correlations; second, because the partial correlations would be taken care of in the factor matrix; and

Table 1  
Intercorrelation Matrix for Twenty Profiles and Film Footage

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Footage		33	41	49	30	29	50	06	02	43	-05	-11	-07	-03	-04	-01	-13	-23	-21	-22	-23
2. I am Learning—1st Showing	33		65	89	90	48	91	37	55	89	33	20	27	26	22	30	12	02	09	05	04
3. Like-Dislike—1st Showing	41	65		76	59	67	71	58	62	73	03	02	-01	05	04	07	-02	-06	-02	-01	-09
4. Predicted Learning—1st Showing	49	89	76		76	55	91	39	53	89	17	07	14	12	10	17	-01	-11	-06	-07	-12
5. I am Learning—2nd Showing	30	90	59	76		47	86	36	52	85	35	24	25	32	27	35	21	07	14	10	11
6. Like-Dislike—2nd Showing	29	48	67	55	47		58	55	50	58	11	03	07	13	09	16	-01	-02	02	02	-07
7. Predicted Learning—2nd Showing	50	91	71	91	86	58		37	53	94	27	12	19	23	16	28	06	-07	-01	-02	-06
8. Rated End Knowledge—3rd Showing	06	37	58	39	36	55	37		78	50	15	15	12	16	08	15	04	09	12	10	05
9. Clarity	02	55	62	53	52	50	53	78		69	16	15	12	15	12	17	07	11	14	16	05
10. Good-Bad Film—3rd Showing	43	89	73	89	85	58	94	50	69		25	18	20	24	20	29	08	-01	05	03	00
11. Test A Learning	-05	33	03	17	35	11	27	15	16	25		74	82	90	71	78	71	60	59	55	65
12. Test B Learning	-11	20	02	07	24	03	12	15	15	18	74		85	84	77	66	75	62	69	56	72
13. True Questions Learning	-07	27	01	14	25	07	19	12	12	20	82	85		72	70	67	68	54	55	49	59
14. False Questions Learning	-03	26	05	12	32	13	23	16	15	24	90	84	72		76	75	77	66	68	62	72
15. Comparable Questions Learning I	-04	22	04	10	27	09	16	08	12	20	71	77	70	76		72	75	59	52	50	61
16. Comparable Questions Learning II	-01	30	07	17	35	16	28	15	17	29	78	66	67	75	72		71	50	62	57	52
17. Maximum Learning	-13	12	-02	-01	21	-01	06	04	07	08	71	75	68	77	75	71		84	85	80	84
18. Terminal Knowledge, Form A	-23	02	-06	-11	07	-02	-07	09	11	-01	60	62	54	66	59	50	84		85	92	88
19. Terminal Knowledge, Form B	-21	09	-02	-06	14	02	-01	12	14	05	59	69	55	68	52	62	85	85		87	86
20. Terminal Knowledge, True Questions	-22	05	-01	-07	10	02	-02	10	16	03	55	56	49	62	50	57	80	92	87		77
21. Terminal Knowledge, False Questions	-23	04	-09	-12	11	-07	-06	05	05	00	65	72	59	72	61	52	84	88	86	77	

Table 2  
Factor Loadings Before Rotation ( $F_0$ )

Variable	I	II	III	IV	V	VI	VII	$h^2$
1	.156	-.439	-.241	.140	-.334	.153	.094	.438
2	.682	-.573	-.235	.199	.205	-.058	-.077	.939
3	.504	-.668	-.204	-.045	-.188	.114	.029	.793
4	.582	-.700	-.189	.157	-.040	.094	-.123	.915
5	.694	-.493	-.192	.188	.165	-.109	.082	.843
6	.456	-.511	.180	-.192	-.231	-.060	.055	.599
7	.653	-.666	-.265	.188	-.036	-.075	-.012	.983
8	.463	-.397	.426	-.434	.134	-.044	.020	.762
9	.538	-.470	.376	-.258	.266	-.058	-.064	.796
10	.696	-.659	-.119	.094	.092	.026	.034	.952
11	.756	.419	-.272	-.163	.033	-.168	-.119	.891
12	.708	.501	-.165	-.188	.133	.221	.100	.892
13	.680	.429	-.274	-.200	.137	.148	-.173	.832
14	.774	.463	-.180	-.133	-.038	-.100	.099	.885
15	.672	.429	-.213	-.155	-.125	.131	.050	.741
16	.716	.352	-.202	-.107	-.088	-.221	.077	.752
17	.697	.608	.069	.143	-.086	.054	.079	.897
18	.599	.620	.360	.233	-.025	.044	-.135	.948
19	.638	.576	.308	.205	.070	-.078	.136	.905
20	.599	.557	.377	.241	-.079	-.147	-.140	.918
21	.600	.640	.206	.182	.135	.128	.104	.890

Table 3  
Rotated Factor Matrix ( $V = FA$ )

Variable	Test Scores Factor (II')	Rating Scales Factor (III')	Terminal Knowledge Factor a (V')	Terminal Knowledge Factor b (VII')	(I')	(IV')	(VI')
1	-.070	.344	-.093	.006	.062	-.231	.223
2	.106	.808	.436	-.095	-.080	-.175	-.051
3	-.010	.746	-.026	.068	.001	-.019	.177
4	.010	.835	.234	.031	-.140	.190	.133
5	.175	.759	.414	-.037	.078	-.158	-.103
6	.057	.678	-.111	.392	.036	.151	-.019
7	.085	.833	.268	-.013	-.033	-.214	-.031
8	.027	.685	-.022	.423	.029	.469	-.064
9	.020	.788	.199	.340	-.047	.313	-.094
10	.084	.899	.314	.059	.026	-.090	.044
11	.787	.055	.088	-.022	-.121	.259	-.164
12	.780	-.004	.133	.005	.111	.309	.197
13	.709	.013	.101	-.095	-.166	.305	.136
14	.850	.044	.087	.094	.097	.231	-.098
15	.773	-.005	-.018	.079	.043	.224	.146
16	.753	.087	.054	.083	.069	.180	-.206
17	.928	-.063	.273	.300	.090	-.024	.029
18	.864	-.057	.391	.484	-.108	-.095	-.012
19	.838	-.003	.433	.406	.163	-.057	-.138
20	.839	-.010	.367	.524	-.119	-.116	-.192
21	.830	-.091	.427	.280	.134	-.028	.063

Table 4  
Cumulative Transformation Matrix (A)

Variable	I'	II'	III'	IV'	V'	VI'	VII'
I	.001	.735	.579	.089	.303	-.003	.290
II	.018	.625	-.777	.102	.000	.039	.000
III	.034	.022	.223	.056	.162	-.072	.828
IV	.012	.096	.000	-.978	.718	-.028	.000
V	.058	-.244	.102	.150	.605	-.126	-.480
VI	.011	.000	.000	.000	.000	.988	.000
VII	.997	.000	.000	.000	.000	.000	.000

Table 5  
Cosine Matrix (C = AA')

Variable	I'	II'	III'	IV'	V'	VI'	VII'
I'	1.000	.000	.000	.000	.050	.000	.000
II'	.000	1.000	-.080	.003	.148	.000	.348
III'	.000	-.080	1.000	.000	.273	.000	.300
IV'	.000	.003	.000	1.000	-.576	.000	.000
V'	.050	.148	.273	-.576	1.000	-.108	-.068
VI'	.000	.000	.000	.000	-.108	1.000	.000
VII'	.000	.348	.304	.000	-.068	.000	1.000

It is the purpose of the present study to analyze the resulting correlation matrix, to determine whether the domain as tapped by the twenty profiles can be readily reduced to a fewer number of dimensions. It is obvious that the measured learning profiles at least must intercorrelate fairly highly, since they were all based on the same instruments. It was the initial hypothesis of this analysis, therefore, that the whole matrix could be reduced to three factors: a measurement factor, with loadings on the tests; a learning factor, with loadings on both the tests and the estimated teaching effectiveness ratings; and an affective rating factor.

### Results

Table 1 presents the intercorrelation matrix. A centroid analysis was applied, yielding the (unrotated) solution given in Table 2. Rotation yielded the final factor matrix in Table 3. The transformation and cosine matrices are given in Tables 4 and 5.

third, because the corrections were applied not to the entire correlation matrix, but to different sectors of it.

Very early in the rotations, two factors practically orthogonal to each other with large loadings emerged. One included all the test profiles (II'), the other included all the rating profiles (III'). For neither of these, nor for any other factor, was the projection of film footage (a measure of the time base) significant, indicating that these were not simply functions of sequence.

Repeated rotational efforts failed to lead to the identification of any other significant factors. There is a suggestion of overlap between ratings of learning and measured learning in factors V' and VII', but the loadings are too small to warrant any large degree of confidence.

### Summary

This analysis suggests that test-measured learning gains from a film are more or less independent of ratings of learning or ratings of affective qualities of the film. Furthermore, the results indicate that all the rating scales used have a large common-factor variance. Whether viewers rate on an "I am learning" scale, a "Like-Dislike" scale, a

"Good-Bad Film" scale, or any similar scale, much the same function seems to be rated.

The findings of this analysis do not agree completely with the conclusion of the original study that a combined test measure is predicted to a significant extent by ratings on the scale "I am learning." Examination of the correlation matrix before and after corrections for "lag," "drift," and "carry-over" (see footnote 5) shows that correlations be-

tween this rating scale and the test score profile were uniformly stepped up as a result of the corrections. It is possible that inclusion of these corrections would have led more definitely to the establishment of a common learning factor, with saturations both on the tests and on the scales rating learning functions.

*Received October 27, 1952.*

## The Relation of Light Intensity to Accuracy of Depth Perception

A. S. Edwards

*The University of Georgia*

The following experiments were performed to discover the relation of light intensity to the accuracy of visual depth perception.

### Procedure

In the first three experiments two lights were used, one of 50 fc, the other of 12 fc. In the fourth experiment, three intensities of light were used, 12 fc, 50 fc, and 100 fc. With all lights, shadows were eliminated.

**Apparatus.** The apparatus was a modified Howard-Dolman depth perception apparatus. Instead of the upright poles, two blocks of wood were substituted. These blocks were made to hold rectangles of heavy pasteboard on which were pasted gray or colored paper. The size of the stimulus thus presented was  $1\frac{1}{2}$ " by  $2\frac{1}{4}$ ". The headrest was at 20 feet distance and kept the S from seeing the tops of the blocks. A standard gray (Stoelting gray No. 19) was always used. The movable stimulus had one of the Stoelting grays or colors as follows: gray No. 19, blue No. 13, green No. 8, yellow No. 4, or red No. 1. Intensities of lights were checked by means of a new light meter from the Physics Department and also by means of one from the local electric power company. New papers were purchased for the experiments.

**Subjects.** The Ss were students chosen at random from classes in the University. Only those with at least 20/20 vision, or vision corrected to 20/20, were used. The Ortho-Rater<sup>1</sup> was used to check vision. Color blind Ss were eliminated.

**Instructions.** Ss were told, "The object of the experiment is to see how accurately you can judge distance by aligning a movable object with a stationary one. You are to take these strings and pull the right hand object

so that it is exactly even with the stationary gray. The right hand string pulls the movable object forward and the left hand string pulls it backwards. When you think the movable stimulus is exactly even with the stationary one, drop the strings. You may move the right hand object back and forth as much as you want to. Keep your chin on the chin rest."

**Experiments 1, 2, and 3.** The first series was run with 40 Ss with 50-fc light, and with 40 Ss with 12-fc light. In the second series there were 30 Ss who were used as both control and experimental Ss with the 12-fc light first. The third series had 30 Ss used as both control and experimental Ss, but with the 50-fc light used first.

**Experiment 4.** This experiment was carried out under the direction of the author by Mr. E. L. Franklin, a graduate student. There were 50 Ss and a 100-fc light was used in addition to the 12-fc and 50-fc lights. The order of lights was alternated.

### Results

On the average it appears that depth perception is more accurate with the greater amount of illumination. But in the first three experiments, it is only in Experiment 1 that this conclusion is substantiated by a critical ratio that indicated significance at the one per cent level. See Table 1.

In Experiment 4, the averages are better as the light intensity increases. But only the difference between the 12-fc and the 100-fc lights gave a critical ratio that was significant at the one per cent level. See Table 2.

**Individual Results.** It appears that some Ss are more accurate with less rather than with more light. In the first three experiments, the per cent of Ss who did better with 12-fc than with 50-fc light were as follows: (a) Experiment 1, 25 per cent; (b) Experi-

<sup>1</sup> Bausch and Lomb Optical Company. *Standard practice in the administration of the Bausch and Lomb occupational vision test with the Ortho-Rater*, 1944, pp. 1-5.

Table 1

Accuracy of Visual Depth Perception with Two Intensities of Illumination

Note: The standard (stationary) stimulus object was at 50 cm. on the scale. M shows the mean on the scale.

Exp.	Ss	50 fc		12 fc		C.R.
		M cm.	S.D.	M cm.	S.D.	
1	80	52.47	3.72	54.72	3.1	3.1 (1%)
2	30	53.07	3.76	54.25	2.7	1.42
3	30	55.71	4.35	56.85	3.45	1.09

Table 2

Accuracy of Visual Depth Perception with Three Intensities of Illumination

Note: The standard stimulus was at 50 cm. and the M gives the average on this scale.

Exp.	Ss	50 fc		12 fc		100 fc	
		M cm.	S.D.	M cm.	S.D.	M cm.	S.D.
4	50	(1) 56.6	1.92	(2) 56.9	1.4	(3) 55.5	2.88

C.R.: 1-2 = 0.93; 1-3 = 2.29 (5%); 2-3 = 3.18 (1%).

ment 2, 30 per cent; and (c) Experiment 3, 20 per cent.

In Experiment 4, which used three intensities of light, the per cent of judgments that were better with the light of less intensity are as follows: (a) 20 per cent better with 12 fc than with 100 fc; (b) 20 per cent better with 50 fc than with 100 fc; and (c) 22 per cent better with 12 fc than with 50 fc.

In our four experiments, it thus appears that a considerable number of Ss judge depth better with less rather, than with greater intensities of light.

### Summary

1. In terms of averages alone, it appears that Ss were more accurate in depth perception with greater illumination rather than with less.

2. Analyses of our data, however, show that from one-fifth to one-fourth of our Ss were more accurate with less rather than with more intense illumination.

3. It appeared that no one intensity of illumination was optimal for all Ss.

Received November 3, 1952.

## Instrument Reading III: Check Reading of Instrument Groups

W. J. White, M. J. Warrick, and W. F. Grether

*Psychology Branch, Aero Medical Laboratory, Wright Air Development Center,  
Wright-Patterson Air Force Base, Ohio*

Most previous studies (1, 3) of instrument reading have been concerned solely with precise quantitative readings. However, it has been pointed out by Grether (2) that there are many situations where instruments are read only to obtain assurance of a null, normal or desired indication (check reading), or for an indication of the direction in which an instrument pointer deviates from a desired value or position (qualitative reading). Milton, Jones, and Fitts (4) report that the average duration of the visual fixations of a pilot flying on instruments is approximately 0.5 seconds. This suggests that aircraft instruments are typically read in a check reading manner.

This paper summarizes a series of experiments concerning the effect of variation in pointer alignment position, in dial diameter and in pointer design on check and qualitative reading of instrument groups. For more detailed reports of these experiments, the reader is referred to a USAF Memorandum Report by Warrick and Grether (6) and two USAF Technical Reports by White (7, 8).

### Experiment I: Effect of Pointer Alignment Position on Check and Qualitative Reading

A series of studies was carried out to determine the effect of various pointer-alignment positions on the speed and accuracy of check and qualitative reading of a four by four arrangement of 16 instruments as shown in the center panel of Figure 1. In addition to alignment at the 9, 12, 3 and 6 o'clock positions, several mixed alignment positions were studied.

**Apparatus and Procedure.** The simulated panel of 16 instruments was presented by means of projected lantern slides. By the operation of a shutter device the subject controlled the exposure of the projected instruments. For the check reading tests, the subject responded by operating a three-position toggle switch held in his lap. He was instructed to move the switch to the

right if all pointers were aligned, and to the left if any pointers were misaligned. Forty college students were used as subjects in this experiment.

A somewhat more complicated response was required in the measurement of qualitative reading performance. To the right of the subject was a group of 16 three-position toggle switches arranged in the same manner as the 16 simulated instruments. If there was a deviating pointer the subject moved the corresponding switch in a direction to indicate an increase or decrease from the alignment position. Forty aviation cadets and forty experienced instrument pilots were used in the experiment. Each of the eighty subjects served under only one condition of pointer alignment.

The apparatus and procedure used in studying mixed alignment differed from that described above. The stimulus material was an actual mock-up of 16 pointers and dials. Only the 9, 12 and 3 o'clock alignment positions were studied. All pointers in any one row deviated by the same amount from the alignment position, the amount of deviation varying between rows. The exposure shutter was operated by the experimenter, who also set the pointers from the rear of the panel. Twelve subjects, staff members of the Aero Medical Laboratory, were used in this experiment, each subject serving under all experimental conditions.

### Results

The results for the check reading study, Table 1, show that the time taken for check reading appears to be relatively independent of the orientation of the pointer. The aver-

Table 1  
Speed and Accuracy of Check Reading a Group  
of 16 Instruments with Homogeneous  
Pointer Alignment

Note: N = 40 (10 subjects in each group).

Pointer Position	Mean Response Time in Seconds	Per Cent of Responses in Error
9 o'clock	0.66	2.0
12 o'clock	0.69	2.5
3 o'clock	0.68	5.0
6 o'clock	0.63	3.0

Table 2

Speed and Accuracy of Check Reading 16 Instruments with Pointers Aligned Diagonally Around the 9, 12 and 3 o'clock Positions

Note: N = 12.

Pointer Position	Mean Response Time in Seconds	Per Cent of Responses in Error
9 o'clock	1.64	18.8
12 o'clock	1.56	15.6
3 o'clock	1.59	10.6

age times required to scan the panel and operate the lap-held switch were 0.66, 0.69, 0.68 and 0.63 seconds for the pointers in the 9, 12, 3 and 6 o'clock positions respectively. Error frequency likewise did not discriminate between the 9, 12, 3 and 6 o'clock alignment positions. With mixed alignment the response times were more than doubled and the frequency of errors was generally higher as shown in Table 2.

For qualitative reading, the exposure and response time data in Table 3 show that alignment at the 9 o'clock position resulted in more rapid reading of the group of 16 instruments than did alignment at the 12, 3 or 6 o'clock positions. A comparison of the error data shows that the readings at the 9 o'clock alignment position were also more accurate. The results are similar for both pilots and cadets. Mixed alignment conditions caused a considerable increase in response time and errors as shown in Table 4.

### Discussion

The apparent superiority of pointer alignment at the 9 o'clock position over the other

cardinal positions for qualitative reading may have resulted from the relationship between the direction of the required switch response and the direction of the pointer deviation. The instructions to the subject read in part, "... if the reading is too much, flip the switch down, ... if the reading is too little, flip the switch up. ...". The direction of the deviation which corresponds to a judgment of "too little" or "too much" changes with alignment position. Apparently the relationship inherent in the 9 o'clock position was optimal under these instructions.

The results of the experiment on check reading did not indicate the superiority of any one of the four alignment positions. They did demonstrate, however, that pointer alignment, particularly uniform pointer alignment, greatly simplifies the task of reading a group of instruments. The response times found in this experiment using 16 instruments are approximately equal to the time required for reading a single instrument in flight.

### Experiment II: Effect of Pointer Design on Check Reading

The purpose of this experiment was to determine the effect of certain factors in the design of pointers on the speed and accuracy of check reading a group of 16 instruments with horizontal pointer alignment. The superiority of pointer alignment for rapid check reading apparently results from the simple pattern which is disturbed by any deviating pointer. Any malfunction which would cause the pointer to rotate 180 degrees might not be detected because of the minor resultant change in the general alignment pattern. Variations in the

Table 3

Speed and Accuracy of Qualitative Reading of Sixteen Instruments with Homogeneous Pointer Alignment

Note: N = 48 (12 subjects in each group).

Pointer Position	Mean Exposure Time in Seconds		Mean Response Time in Seconds		Per Cent of Responses in Error	
	Pilots	Cadets	Pilots	Cadets	Pilots	Cadets
9 o'clock	0.90	1.17	1.63	1.84	2.91	5.00
12 o'clock	1.15	1.39	2.02	2.17	6.25	12.08
3 o'clock	1.30	1.83	2.32	2.34	23.33	20.83
6 o'clock	1.62	1.46	1.89	2.29	10.83	9.58

Table 4  
Speed and Accuracy of Qualitative Reading 16 Instruments with Pointers Aligned Diagonally Around the 9, 12 and 3 o'clock Positions  
Note: N = 12.

Pointer Position	Mean Response Time in Seconds	Per Cent of Responses in Error
9 o'clock	4.51	15.8
12 o'clock	4.36	22.6
3 o'clock	4.78	14.9

design of the base of the experimental pointers in the present study were intended to increase the change in pattern resulting from a 180-degree deviation.

*Apparatus and Procedure.* The stimulus material and method of presentation and procedures were very similar to those used in the preceding experiments in this report. The five pointer designs studied in this experiment are shown in Figure 1. All pointers had about the same physical dimensions. They were 1 1/8" long and 3/32" wide at the center. Five film strips, one for each

pointer design, were made of a panel containing 16 simulated 1 3/4" instruments. Forty instrument pilots and an equal number of cadets served as subjects; each served under only one condition of pointer design.

Results

The results in terms of speed and accuracy of check reading as a function of pointer design are summarized in Table 5. In the first two columns of this table are the per cent of readings in error, all errors being combined for pilots and cadets. In the next two columns are the per cent of undetected 180-degree deviations. For both subject groups the number of undetected deviations was greatest for design D. For the pilot group only the largest difference, that between designs D and E, is significant at the 5% level of confidence. For the cadet group none of the differences are significant at the 5% level of confidence. This suggests that the experimental pointers (designs A, B, C and E), when rotated 180 degrees from alignment position, were little if any more effective in pro-

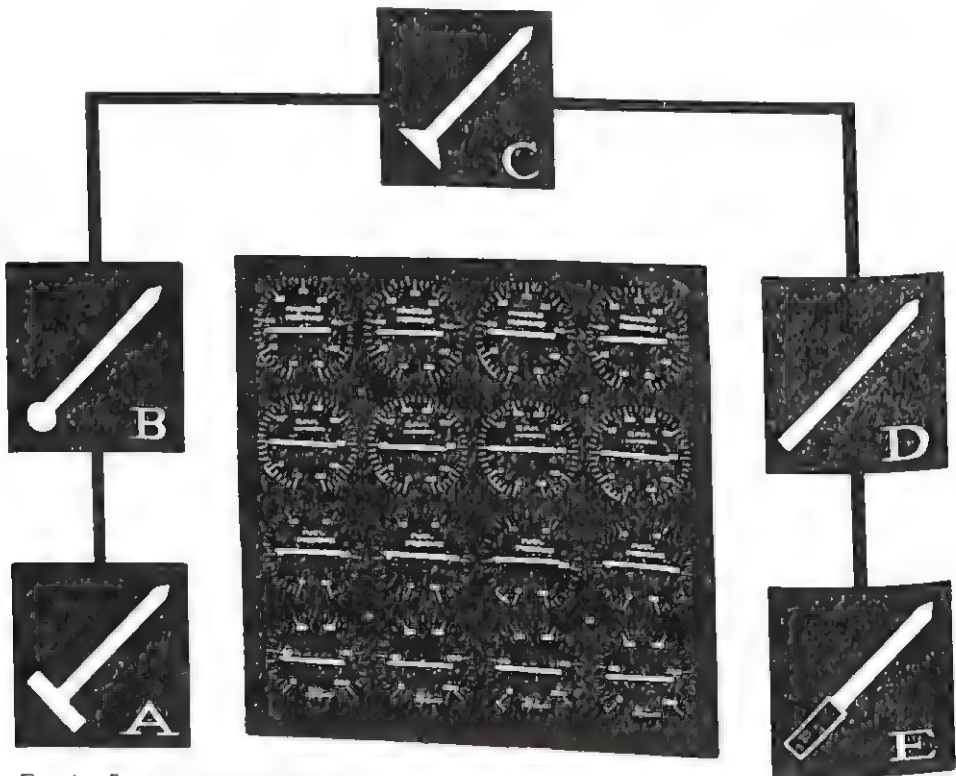


FIG. 1. Instrument panel arrangement used in all check reading experiments and pointer designs used in Experiment II.

Table 5

The Effect of Pointer Design on the Speed and Accuracy of Check Reading 16 Instruments with Homogeneous Pointer Alignment

Note: N = 90 (10 subjects in each sub-group).

Pointer Design	Per Cent of Readings in Error (N = 200 Trials)		Per Cent of Undetected 180° Deviations (N = 40 Trials)		Per Cent Misc. Errors (N = 160 Trials)		Mean Response Time in Seconds	
	Pilots	Cadets	Pilots	Cadets	Pilots	Cadets	Pilots	Cadets
A	14.00	11.00	45.00	37.50	6.25	4.37	2.00	2.11
B	16.50	10.50	47.50	35.00	8.75	4.37	1.62	1.91
C	15.50	15.5	47.50	27.50	7.50	12.50	1.80	2.05
D	14.50	12.5	62.50	47.50	2.50	4.37	1.75	1.90
E	21.50	—	39.00	—	16.00	—	2.80	—

viding cues for detection of misalignment than was the standard pointer (design D).

If we compare the data for miscellaneous errors in Table 5, a somewhat different picture is obtained. For the pilot group the error percentage was lower for the standard pointer, design D, than for other designs with the differences significant at the 5% or greater level of confidence. The results for the cadet group do not support this superiority of the design D. The mean response time values in the last two columns of Table 5 show no significant differences between pointer designs.

#### Discussion

The failure of the experimental pointers to reduce substantially the 180-degree type error may, perhaps, be attributed to a configurational change brought about when the pointer overlays the dial numerals and graduations. This reduces the visibility of the expansions

of the pointer base. This is especially likely when the panel is scanned rapidly as in check reading.

#### Experiment III: Ocular Movements in Relation to Dial Diameter

The pioneering work of Paterson and Tinker (5) on typographical factors which influence reading suggests that there might be an optimal dial diameter for instrument reading purposes. The optimal dial diameter should be such that the resulting instrument panel size could be scanned accurately, rapidly and with a minimum number of eye fixations. The most suitable approach to this problem appeared to be through the combined measurements of manual responses and of eye movements while check reading a panel of instruments.

*Apparatus and Procedure.* For this experiment three panels were used. The instrument diame-

Table 6

Eye Movements, Response Times and Errors in Check Reading a Group of 16 Instruments

Note: N = 6.

Type of Measure	Dial Diameter		
	1 inch	1½ inches	2¼ inches
Mean number of fixations per trial	2.8	2.6	2.9
Mean time per fixation	0.29	0.26	0.26
Total scanning time from preliminary to final fixation	0.82	0.72	0.75
Per cent deviating pointer trials on which deviating pointer was fixated	72	71	81
Response time (sec.)	0.76	0.72	0.73
Per cent errors	6	4	12

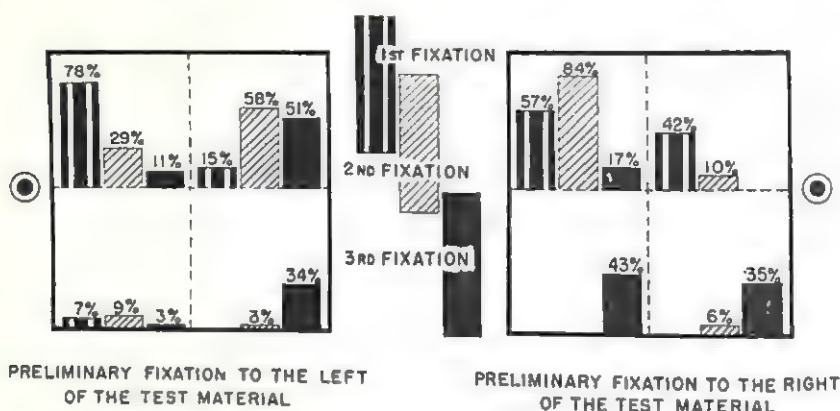


FIG. 2. Distribution of first, second and third eye fixations in check reading a panel of 16 instruments.

ters were 1, 1 $\frac{1}{4}$ , and 2 $\frac{1}{4}$  inches. In all cases the spaces between the margins of the instrument were  $\frac{1}{8}$  inch. These panels were check read using both eyes. The movements of the right eye were recorded by the corneal reflection technique. An American Optical Company Ophthalmograph was modified so that exposures could be made with either moving or stationary film. The moving film exposures provided data on the number of fixations, and the lateral position and duration of fixations. The stationary film exposures provided additional data on the position of fixation in both the vertical and lateral dimensions, but did not record the duration of fixations.

On each of 60 test trials the subject was required to fixate a spot to the left of a screen covering the panel. When the screen was raised the subject scanned the panel as needed and then fixated on a spot to the right of the screen. He then signalled his judgment as to whether or not a pointer was deviated from the aligned position. For three of the six subjects the initial and final fixation points were reversed. When the film record of the subject's eye movements was projected against a replica of the corresponding instrument panel, the initial and final fixation points provided references for the analysis of the film records.

### Results

The major results obtained from the recording of eye movements are summarized in Table 6, which gives check reading time, response time and error data in addition to the various measures of visual performance. Although not statistically significant, the results are quite consistent in showing slightly superior performance with the 1 $\frac{3}{4}$ -inch dial size.

Analysis of the location of the 1st, 2nd and 3rd fixations, in relation to the panel arrange-

ment, are shown in Figure 2. The major proportion of the first fixations were in the upper left quadrant of the instrument panel. This was true even when the preparatory fixation was at the right of the panel. Surprisingly few fixations were made on the lower two rows of dials, and those that did fall in this area were usually the last fixation in the scanning sequence. Analysis of the errors committed shows that 22 per cent more errors were made when the deviating pointer was in the lower half of the panel. In the fourth row of Table 6 is shown the per cent of deviating pointer trials in which the deviating instrument was fixated.

### Conclusions

The following conclusions are suggested by the results of the foregoing experiments:

1. The pointer alignment position (9, 12, 3 or 6 o'clock) has little effect on simple check reading involving the mere detection of a pointer deviation.
2. Uniform horizontal or vertical pointer alignment facilitates instrument-check reading.
3. Pointer alignment at the 9 o'clock position is optimum for qualitative reading as defined by the perceptual judgments and manual responses required in these experiments.
4. Pointer modifications of the type employed in the present experiment do not accomplish a satisfactory reduction in difficulty of detecting 180-degree deviation errors.

5. A  $1\frac{3}{4}$ -inch instrument dial appears superior to either a larger or smaller dial in terms of the number and duration of eye fixations while check reading.

Received August 25, 1952.

### References

1. Fitts, P. M. *Engineering psychology and equipment design*. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: John Wiley & Sons, 1951, 1287-1340.
2. Grether, W. F. Instrument reading. I. The design of long-scale indicators for speed and accuracy of quantitative readings. *J. appl. Psychol.*, 1949, 33, 363-372.
3. Kappauf, W. E. *Studies pertaining to the design of visual displays for aircraft instruments, computers, maps, charts, tables and graphs: A review of the literature*. Engineering Division, Air Materiel Command, Dayton, Ohio, USAF Technical Report No. 5765, April, 1949.
4. Milton, J. L., Jones, R. E. and Fitts, P. M. *Eye fixations of aircraft pilots, II. Frequency, duration and sequence of fixations when flying the USAF instrument low approach system*. Engineering Division, Air Materiel Command, Dayton, Ohio, USAF Technical Report No. 5839, October, 1949.
5. Paterson, D. G. and Tinker, M. A. *How to make type readable*. New York: Harper & Bros., 1940 (out of print).
6. Warrick, M. J. and Grether, W. F. *Effect of pointer alignment on check reading of engine instrument panels*. Engineering Division, Air Materiel Command, Dayton, Ohio, USAF Memorandum Report No. MCREXD-694-17, June, 1948.
7. White, W. J. *Effect of dial diameter on ocular movements and speed and accuracy of check reading groups of simulated engine instruments*. Engineering Division, Air Materiel Command, Dayton, Ohio, USAF Technical Report No. 5826, June, 1949.
8. White, W. J. *Effect of pointer design and pointer alignment on speed and accuracy of reading groups of simulated engine instruments*. Engineering Division, Air Materiel Command, Dayton, Ohio, USAF Technical Report No. 6014, July, 1950.

## Dimensional Analysis of Motion: VI. The Component Movements of Assembly Motions<sup>1</sup>

Robert Smader and Karl U. Smith

*The University of Wisconsin*

### Methods

The application of electronic methods to investigation of human motions (1, 4, 5) provides the means of systematic experimental study of long-standing problems of motor coordination. Such electronic techniques, which have been used previously in studying panel-control operations and closely related performances, are used in this study to analyze the component movements in assembly motions. Data are presented bearing upon the relative efficiency of the component movements of assembly motions and upon the effects of practice on such movements.

Two methodological principles are demanded for the scientific investigation of complex human motions. First, the basic dimensions of motion must be subject to control and quantitative specification. Second, the component movements of the pattern must be segregated and measured separately. This second requirement for motion study must be achieved with an economy sufficient to permit replication of the observation of the same motion pattern throughout the course of an experiment.

Control and specification of the various dimensions of movement, especially the bodily and space dimensions of movement, are secured through preplanning of performance situations in order to permit quantitative variation in the direction, plane, magnitude, bodily orientation, and other aspects of motion. Measurement of the component movements in assembly motions may be achieved by means of electronic motion analysis techniques, to be described, which provide, for the first time, methods adequate for the experimental analysis of the separate movements of skilled performances.

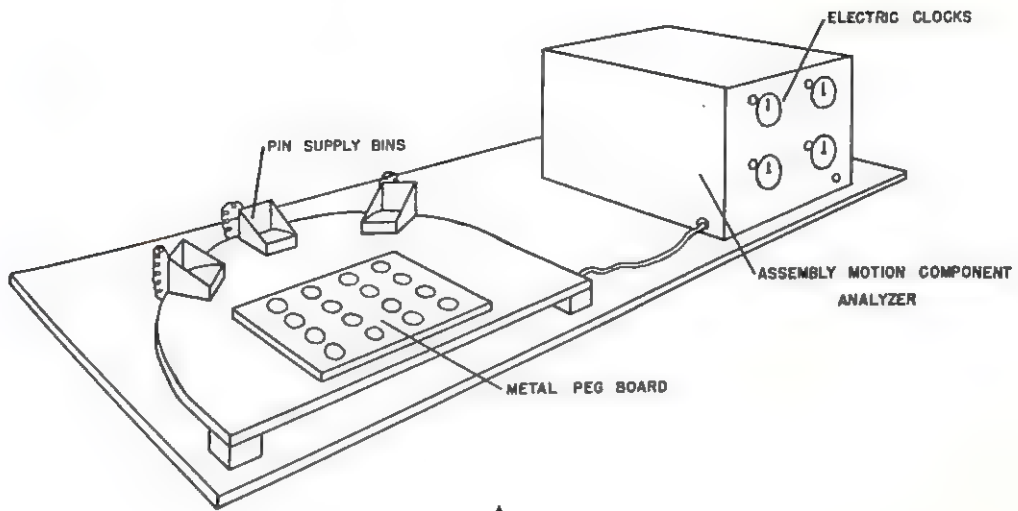
<sup>1</sup> This research has been supported by funds voted by the Legislature of the State of Wisconsin, and assigned by the Graduate School Research Committee, the University of Wisconsin.

Figure 1A illustrates the main parts of the apparatus used for electronic registration of the duration of the component movements in assembly motions. The task performed by the subject in this situation, as illustrated by Figure 1B, consists of four major movement components: (a) an initial reaching or travel movement that carries the hand and arm to the bins containing pins used as assembly objects; (b) a grasping movement in securing one of the pins; (c) a return or loaded travel movement that carries the pin to the assembly plate; and (d) a positioning motion of inserting the pin in the assembly plate.

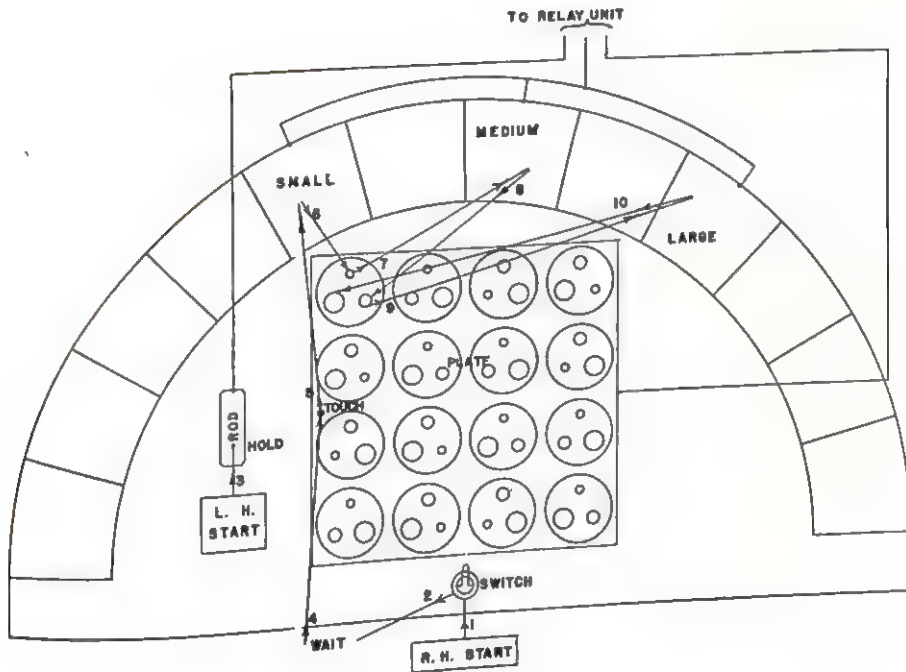
The preplanned performance situation used here is designed to provide for systematic variation in different space dimensions of component movements of the assembly task (Figure 1B). Assembly bins and the assembly plate are arranged generally on a work table measuring 35 in. by 20 in. The metallic assembly bins, which are fitted with metal bases, are 8 in. long, 4 in. wide, and 4 in. high. The floor or base of each bin is curved to facilitate grasping of the pins located inside. Six such bins may be used for pins or collars of different sizes. The distance and arrangement of these bins with respect to the assembly plate may be changed to vary the magnitude, direction and plane of component travel and grasping movements.

The assembly plate (Figure 1B) is of aluminum  $\frac{1}{2}$  in. thick,  $10\frac{1}{2}$  in. long and  $10\frac{1}{2}$  in. wide, drilled with 16 large apertures,  $1\frac{1}{16}$  in. in diameter. Fitted into these apertures are discs which are drilled with three holes of different sizes, i.e.,  $\frac{1}{4}$  in.,  $\frac{3}{8}$  in. and  $\frac{1}{2}$  in. The discs may be locked into three different positions on the assembly plate. Three sizes of pins, all approximately one inch long but of diameters to fit the three sizes of holes, are provided in three bins. Variations in size of object grasped, serial order of placement of the pins, and other variations in the dimensional aspects of the assembly placing movement are made by means of the different placements of the bins and positions of the holes on the assembly plate.

The subject's movements in this situation are recorded by means of an electronic motion analyzer that automatically segregates and measures separately in 0.01 seconds the duration of each of the four basic component movements in the task. Figure 1A illustrates how the clocks and circuits of the analyzer are arranged with respect



A



B

FIG. 1. A diagrams the general layout of the motion analyzer circuits, the time clocks, the assembly plate, and the bins. The bins are made with curved bottoms in order to facilitate removal of the pins. B shows the surface detail of this same setup. The letters "L.H." and "R.H." in this diagram refer to left hand and right hand respectively. The sequence of movement with the right hand is indicated by number. The size of pins located in the bins is indicated in relation to the position of the bin.

to the work area. The basic principle of operation of the electronic analyzer is that the subject acts as a key in the circuit and thus sequentially activates different relays and clocks<sup>2</sup> during the different stages of the assembly motion. The current level used in the subject's side of the circuit is below cutaneous threshold level, so that the operator has no sensory appreciation of the fact that he is serving as a conductor.

Figure 1B shows how the device performs in actual operation. Upon starting a cycle of assembly motion, the subject initially touches the assembly plate. This contact with the plate starts the first of the four precision time-clocks, the initial travel time-clock. Upon coming into contact with the pins in one of the assembly bins, this first clock stops and the second, the grasping time-clock, starts. When the contact between the pin and the bin is broken, the grasping time-clock ceases to run and the loaded-travel time-clock starts to run. This third clock stops when the pin is brought into contact with the assembly plate, and the fourth clock, the assembly positioning time-clock, starts to run. This fourth

<sup>2</sup> Model S-1, Standard Electric Time Company, Springfield, Mass.

clock in turn stops running when electrical contact with the assembly plate is broken and a second cycle of motion is initiated. The different times registered on each clock may be read after one cycle of motion or they may be allowed to summate over several cycles and then be read.

In the experiment to be described, 46 right-handed college students were used as subjects. These subjects were given standard instructions to fill the assembly plate with pins according to a defined sequence, which was kept constant for all subjects. One complete filling of the plate constituted a trial. Subjects performed two trials per day for each of three consecutive days. All scores analyzed were the mean durations of each component movement for each day.

In addition to the running of the test trials, as just described, a calibration trial was run before the first test trial of the second and third days. In these calibration trials, the two manipulation time clocks and the two travel time clocks were electrically connected so that the discrepancy between two clocks in measuring the same elapsed time could be determined.

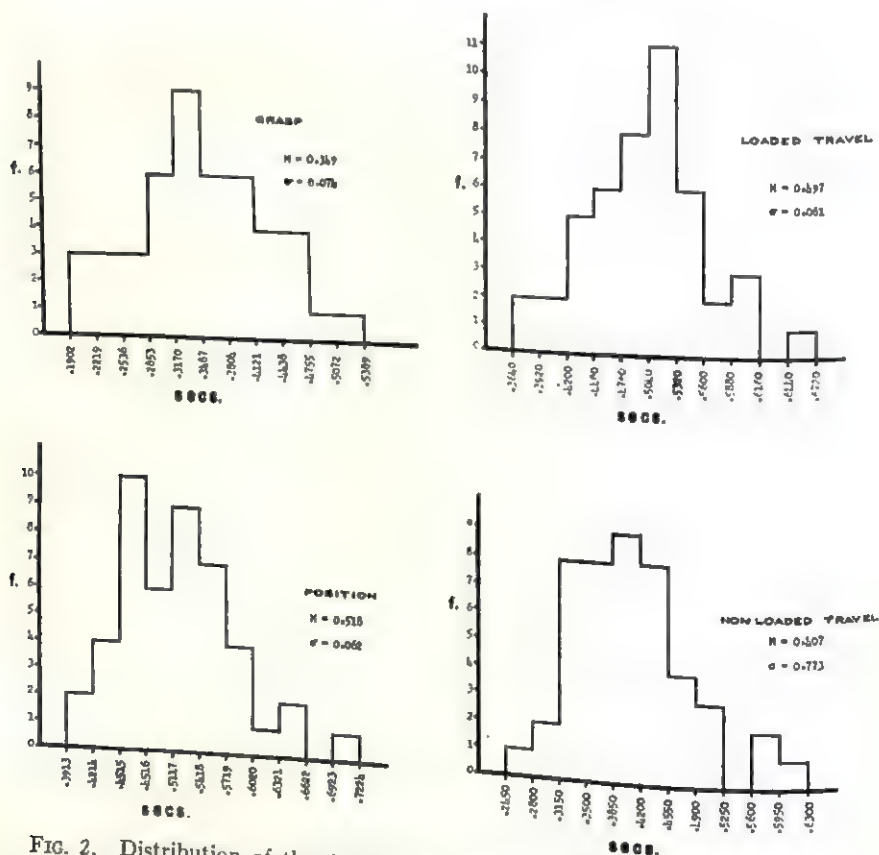
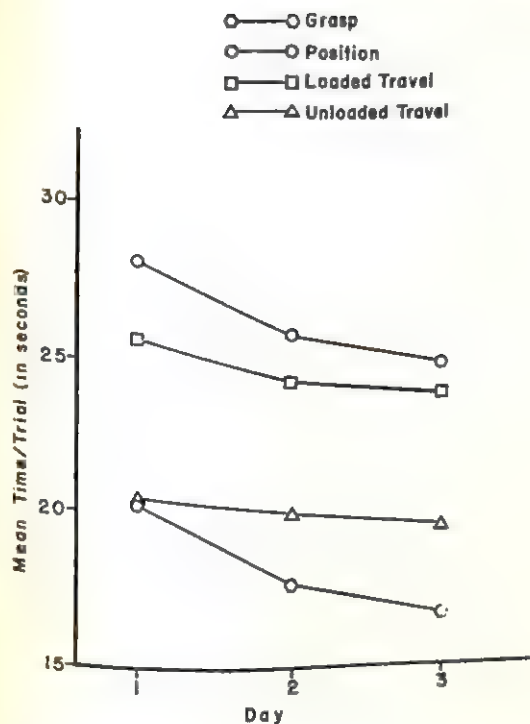
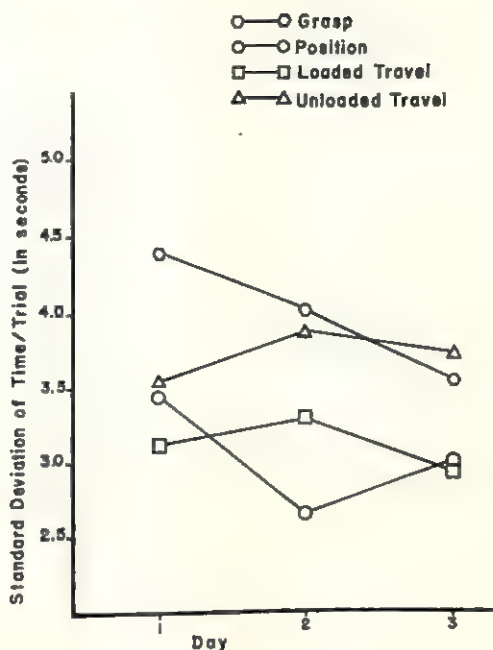


FIG. 2. Distribution of the durations of different component movements in relation to subjects. The figures for the manipulative components are given to the left and those for the travel components to the right.



A



B

FIG. 3. The change in the duration (A) and in the variability (B) of the four different component movements as a function of practice. Generally, the two travel movements show less change with practice than that observed in the grasping and positioning movements.

### Results

The results of this experiment will be presented with respect to the following main points: (1) distribution of time values for different component movements in the assembly task; (2) the effects of practice on different component movements in the task as well as on their interrelation; and (3) the reliability of measurements of different movement components.

Figure 2 shows the distribution of movement times of the four component movements of the assembly task on the third day of practice. The data presented in this figure represent the durations of individual movements for each part of the assembly task. The shortest movement component giving the shortest times is the grasping movement. The assembly positioning movement gives the longest durations. All four movement components show distributions that are slightly skewed positively.

Figures 3A and 3B describe the change in the duration of the component movements of the assembly task as a function of practice. Figure 3A gives the learning curves based on the mean duration of each of the component movements on each day. Figure 3B shows the change in variability of each component movement as a function of practice.

Three of the four movement components in the task show a highly significant change in duration as a result of practice over three days. The greatest change, about 18 per cent, is found for the grasping component of the task. The other manipulative component of the task, the positioning movement, gives an 11 per cent change with practice. Both of these manipulations show a greater change with practice than does either of the travel movements. One of the latter, the unloaded travel movement, shows a statistically significant change with practice only at the 0.05 point. Tests of significance of the differences

between day 1 and day 3, based on "t," indicate that changes occurring in learning for the three other components of movement are significant at the 0.001 point. On each day of practice, the differences between the mean durations of the two manipulative components, grasping and positioning, and between the two travel movements are statistically significant at the 0.001 level.

The variability of different component movements does not change uniformly as a function of practice. The variability of the unloaded travel movement decreases with practice, but not significantly. In contrast, the variability of the grasping motion is reduced sharply with practice. The other manipulative movement, positioning, has its

standard deviation changed about 15 per cent with practice. The variability of the loaded travel movement is hardly changed at all during the three days.

The present method of isolating electronically the different components of movement in assembly has provided the means of studying the interrelation between separate movements as a function of practice. In order to conduct such investigations, the individual mean scores for each component of motion are obtained for each day of practice and correlated. Thus, six correlations between pairs of the four component scores are obtained for each day of the study.

The correlation values just described are plotted in Figure 4 as a function of days of

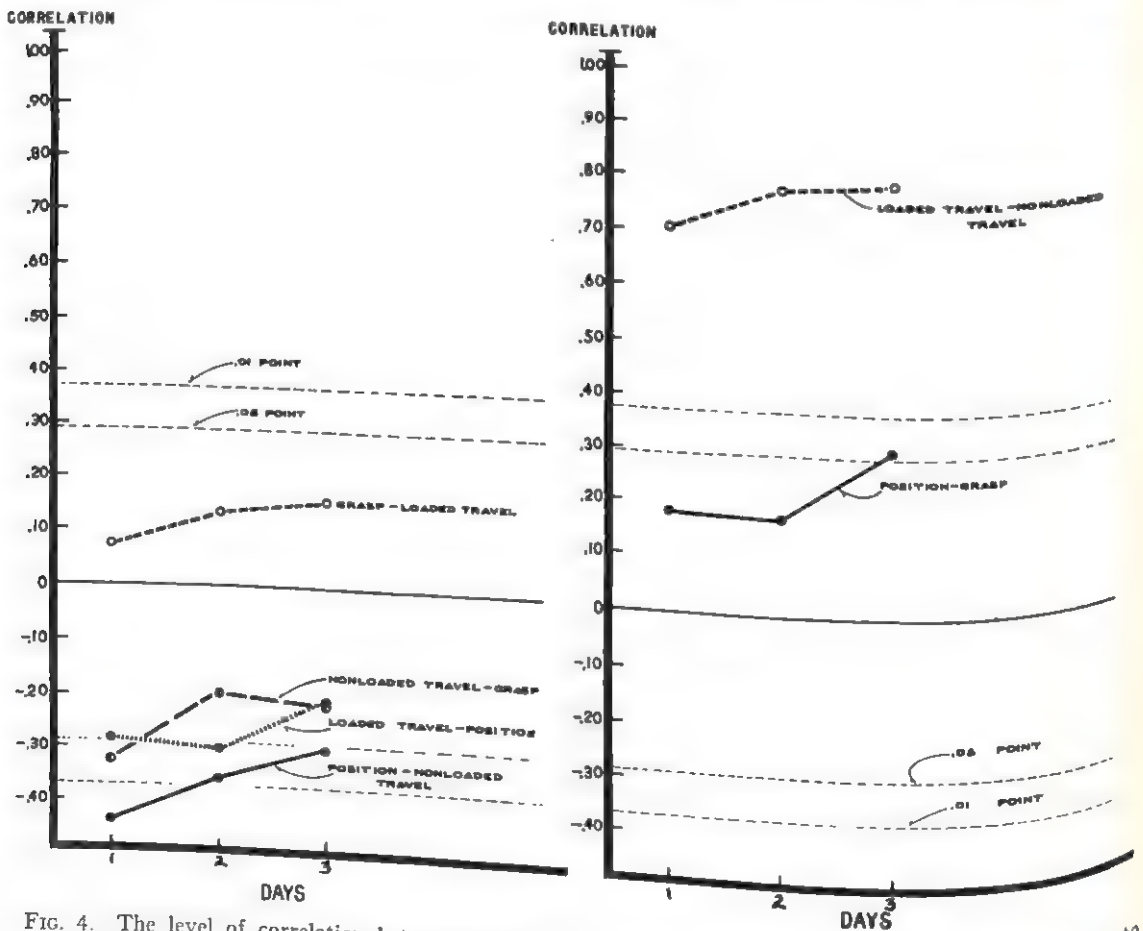


FIG. 4. The level of correlation between different component movements during practice. The plot to the left gives the curves for pairs of movements that are adjacent to one another in the motion pattern. The plot to the right gives the curves for pairs of movements that are non-adjacent in the cycle. The curves, only those for the correlations between loaded and nonloaded travel and between positioning and nonloaded travel are statistically significant for all three days of practice.

Table 1

Coefficients of Stability of the Component Movement Times

Type of Component	Component	Day 1 vs. Day 2	Day 2 vs. Day 3
		Test-retest Reliability Coefficient	Test-retest Reliability Coefficient
Manipulative	Position	+0.79	+0.86
	Grasp	+0.79	+0.91
Travel	Unloaded	+0.83	+0.88
	Loaded	+0.80	+0.91

practice. In addition, the broken lines on Figure 4 indicate the 5% and 1% levels at which these correlations may be considered to differ significantly from zero. The correlation functions which give values consistently significant at the 5% level are those representing the relations between loaded and unloaded travel movements and between position and unloaded travel movements. Although it might be expected that the position and grasp movements would correlate highly with one another, these two movements in fact show no significant relation with one another.

The main point to be observed in Figure 4 is the change in the correlation between component movements in relation to practice period. It will be observed that none of the correlations between different pairs of movements is significantly altered during the three days of practice. In other words, the interrelation of movement components does not vary during learning and improvement in performance in the over-all task. If the correlation between component movements is thought of as a measure of integration between these movements, then the results show that learning has little or no effect on the integration of separate movements in assembly motions.

Inasmuch as the present experimental techniques represent quite complex procedures for automatic measurement of behavior, it is of importance to examine the temporal consistency of scores obtained. It will be remembered that each subject performed two trials on each of the three days of practice, and that four scores, representing different component movements, were obtained on each

trial. To obtain correlation coefficients between days for each of the four component movements in the task, a mean score for each subject for each type of movement was obtained and these values for different days were then correlated with one another. Table 1 summarizes the values obtained. The table shows that a high degree of consistency is found for the component movement times in the assembly task. It should be noted again that the figures given in Table 1 are based upon only two trials of performance per day and that the figures represent consistency between days. These figures in Table 1 show that the low intercorrelations between component movements, as given in Figure 4, cannot be accounted for in terms of unreliability of the individual movement times.

### Summary

This study describes electronic techniques which make possible detailed experimental study of the component movements in assembly motions. Besides indicating the nature and value of these methods for analysis of industrial motions, the present paper describes experimental results bearing upon the characteristics of component movements in an assembly task, and upon the effects of practice on the duration and relation between these movements.

By means of the electronic techniques described, the component movements of travel, grasp, loaded travel and positioning in a unimanual assembly task may be isolated and their durations measured in hundredths of seconds. Distributions of the component movement times for forty-six college student operators are approximately normal.

Practice does not affect uniformly different types of movement in the assembly task. The efficiency of the two manipulative movements, positioning and grasping, are changed the most by practice. In contrast, the travel motions in the task, especially the unloaded travel component, show very little change with practice. Similar results are found for measures of variability of response during practice.

It has been observed also that practice does not alter significantly the correlation between

component movements in the assembly task. Early in learning the correlations between the different component movements of manipulation and travel are near zero. As practice continues, the values of these correlations do not change significantly. At the start of learning, a correlation of about  $+0.07$  is found between the two travel movements in the task. With practice, this correlation increases slightly but not significantly.

The results on the effects of practice, as just described, are equivalent to experimental findings obtained in previous studies of learning and component movements in tracking behavior (2, 3) and panel control motions (1, 4, 5, 6). In all of these tasks, the human body seems to act like a many-channeled system in the production of discrete unrelated movements for a given task, the integration of which is not altered materially by continued practice. The findings described here, along with similar results found for other motion patterns, point toward the lack of significance of learning concepts in understanding the details of movement coordination in skill. The same results, however, point up the great significance of component and dimensional motion analysis in dealing practically and theoretically with problems of skilled motion and the instrumental relations of such motion.

The electronic methods of motion analysis described in this report provide, for the first time, economical methods of obtaining reliable measures of movement components in assembly skills. These methods, along with procedures of preplanning and quantitative control of the dimensions of motion in work, lay the foundation of scientific study of motion in terms of modern experimental designs.

Received October 6, 1952.

### References

1. Davis, R., Wehrkamp, R., and Smith, K. U. Dimensional analysis of motion: I. Effects of laterality and movement direction. *J. appl. Psychol.*, 1951, 35, 363-366.
2. Lincoln, R. S., Simon, J., and DeCrow, T. W. Effects of practice upon different component movements in tracking. *Per. and Mot. Skills Res. Exch.* (In press).
3. Lincoln, R. S., and Smith, K. U. Systematic analysis of factors determining accuracy in visual tracking. *Science*, 1952, 116, 183-187.
4. Rubin, G., and Smith, K. U. Learning and integration of component movements in a pattern of motion. *J. exper. Psychol.* (In press).
5. Smith, K. U. and Wehrkamp, R. A universal motion analyzer applied to psychomotor performance. *Science*, 1951, 113, 242-244.
6. Wehrkamp, R. and Smith, K. U. Dimensional analysis of motion: II. Travel distance effects. *J. appl. Psychol.*, 1952, 36, 201-206.

## The F minus K Index on the MMPI

Angus G. MacLean, Arthur T. Tait, and Calvin D. Catterall

*California Test Bureau, Los Angeles*

Hunt (2) compared F-K raw scores for honest and dissembled MMPI profiles from the same subjects and found that an F-K cutting score of  $-11$  and below was fairly effective in spotting fake-good cases, but also picked up too many supposedly honest cases. Gough (1) reported the distribution of F-K scores for a group of 691 adult normals, the middle 80% of whom obtained approximately normally distributed scores ranging from  $-22$  to  $+11$  with a mean of about  $-9$ .

The present data consist of a sample of 100 candidates for nursing randomly drawn from a large number of similar cases in the Los Angeles and San Francisco Metropolitan areas. These candidates are almost without exception female, are finishing the 12th grade or are in their first year of college; if not attending college, they range from 18 to 22 years of age with a modal age of 18 to 19. These cases provide data worth noting since they are in a selection situation; furthermore, the balance of their MMPI profiles suggests that they are motivated to present themselves in the best light. Few T-scores exceed 60 on any clinical scale and virtually none exceed 70.

The F, K, and L scores for these subjects are distributed as follows: Actual distribution of F scores:  $T = 50$ : 64 cases;  $T = 53$ : 17 cases;  $T = 55-59$ : 8 cases;  $T = 60-64$ : 10 cases; and  $T = 73$ : 1 case.

A mean and standard deviation should not be calculated for such a distribution.

The distribution of K scores had a mean of 57.9 and a standard deviation of 8.5. The total range was from 30 to 79, and the middle 80% of cases lay between 46 and 69.

The L scale had a mean of 72.0 and a standard deviation of 8.1. The total range was from 50 to 84, and the middle 80% of cases lay between 61 and 82. The scores were slightly negatively skewed, and it seems possible that these candidates, while defensive on the F and K items, felt that frankness on the

L items would not be damaging, or even that they might be "catch" items.

The distribution of raw scores, somewhat smoothed, is shown in Table 1.

It is suggested that a cutting score of  $-17$  might be tentatively adopted: this would include 31% of the nurse candidates but only about 12% of the "Adult Normal" cases discussed by Gough. A cutting score of  $-14$  would include 50% of nurse candidates and

Table 1

F-K Raw Score Distribution on MMPI Based on a Sample of 100 Candidates for Nursing

Sten Score (1 sten = 0.5 $\sigma$ )	Raw Score (Lower Limits)	% in Category (Smoothed)	Cumulative %
9	-25 up	2.3	2.3
8	-22	4.4	6.7
7	-20	9.2	15.9
6	-17	15.0	30.9
5	-14	19.1	50.0
4	-11	19.1	69.1
3	-9	15.0	84.1
2	-7	9.2	93.3
1	-2	4.4	97.7
0	-1, 0, or positive	2.3	100.0

about 22% of Gough's "Adult Normals." It may be assumed, lacking information to the contrary, that the "Adult Normals" consisted merely of non-pathological cases, and included quite a wide range of cases from the moderately self-critical and fairly honest to the highly ego-defensive. About 1% of Gough's cases obtained F-K scores of  $-22$  or more. Thus, if a cut-off is established at  $-17$ , about 12 or 15% of Gough's cases will be included, it is true, but it may well be supposed that these cases are not so much unusually well-adjusted, as those with a tendency to dissemble even in a non-selection situation.

In view of the above data, the following ranges for the F-K index are suggested:

1. Positive Raw Score, especially if greater than + 2: malingerers or very self-critical and unusually honest subjects.
2. Raw Score from zero to - 10: Normal Area.
3. Raw Score from - 11 to - 16: Doubtful Area.
4. Negative Raw Score greater than - 17: Probable "Fake-good" Area, especially when the F-K difference lies in the twenties.

Optimal cut, as always, refers to a specified pair of populations, and caution must be used in generalizing to dissimilar groups and situations. For example, in another study we found the F-K raw-score difference to range from - 25 to + 30, with a middle-80% range from - 16 to + 9, a mean of - 5 and a median of - 6. The SD for this group is

approximately 10 raw-score points. We can only assume that some members of this group desired to appear in a good light, while others were either malingerers or unusually self-critical. This would suggest regarding small positive scores as indicative of frankness and insight, while the suspicion of malingerers would pertain to positive scores in the neighborhood of + 15 or more.

*Received February 11, 1953.*

*Early publication.*

### References

1. Gough, H. G. The F minus K dissimulation index for the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1950, 14, 408-413.
2. Hunt, H. F. The effect of deliberate deception on Minnesota Multiphasic Personality Inventory performance. *J. consult. Psychol.*, 1948, 12, 396-402.

## Psychological and Personal History Data Related to Accident Records of Commercial Truck Drivers \*

James W. Parker, Jr.

*Tufts College*

This study is one phase of a larger over-all research project being carried out in the Department of Psychology at North Carolina State College, Raleigh. The general purpose is to improve the method of selecting truck drivers to be employed by a large East Coast trucking concern.

Criterion data for studies such as this have long been a subject of controversy. Up to the present time the data used have been the accident records. The manner in which these data have been used has been varied, but no really "new" idea in criterion data has been devised. It is basically the accident record, therefore, which was used in this study, and the records for the various subjects were equated on the basis of the number of miles driven by each.

The subjects were 104 drivers who were still employed by the company on a certain cut-off date, and who had been trained at the Driver Training School operated by the North Carolina State College. All subjects had been tested at the training school by an examiner trained in psychological testing, and under standardized conditions. In most cases the test results were not used in the employment procedure by the company. Some subjects did not have a complete report of their test performance in their files, and this necessitated using varying sample sizes for the different variables.

The factors or variables with which the accident records are compared are divided into two categories: (a) psychological test data and (b) personal history data which were derived from the application blank filled out by the employee at the time of his em-

ployment by the company. The tests used were some of the better-known paper-pencil tests of intelligence, mechanical comprehension, personality, and vocational interest. The results of visual screening with the Bausch and Lomb Ortho-Rater were also included. Personal history data included such things as age, education, marital status, blood pressure, etc.

Two sources were used to arrive at the criterion data: the number of accidents a man had incurred during a certain period of time, and the total number of actual miles driven during this same period of time. The accident rate was then put on the basis of the number of accidents per 5,000 miles. This was done for two types of accidents, those classified preventable and non-preventable by the safety department of the trucking company.

### Procedure

The distributions of scores for the entire sample of 104 drivers on the psychological test data and the personal history data were divided into two groups, accident group and non-accident group, with respect to each of the two criteria, preventable and non-preventable accidents per 5,000 miles. The accident group for each of the criteria was further divided into upper and lower halves, excluding the accident-free group. Figures 1 and 2 show the distributions of preventable and non-preventable accidents.

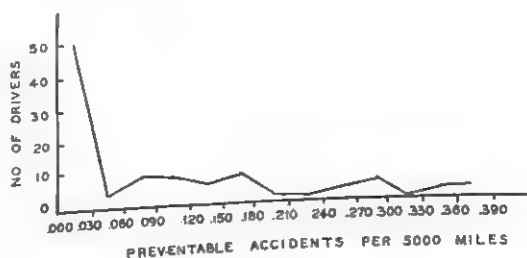


FIG. 1. Distribution of preventable accidents.

\* This study is a condensation of a thesis submitted to the North Carolina State College of Agriculture and Engineering of the University of North Carolina, Raleigh, in partial fulfillment of the requirements for the degree of Master of Science in Industrial Psychology. The author wishes to express appreciation to Dr. Dannie J. Moffie under whose direction this study evolved.

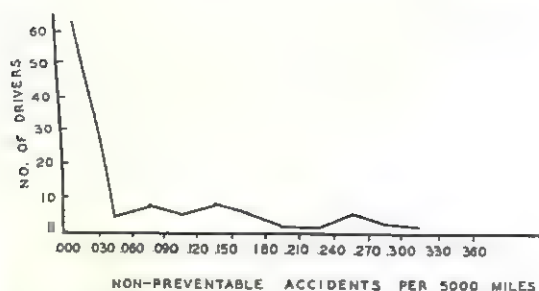


FIG. 2. Distribution of non-preventable accidents

"Student's"  $t$ -ratio (2) was run between the means of the groups for each of the variables as follows: between the accident and non-accident groups, and between the upper and lower halves of the accident group.

The  $F$ -ratio test of differences was run between the variances of the groups mentioned above, and in like manner. This was done in order to determine if the distributions being compared by means of the  $t$ -ratio had significantly different dispersion within the distributions. In any analysis where the  $F$ -ratio was significant at the 5 per cent level of confi-

dence or better, and the significance level of the  $t$ -ratio was critical, the significance levels of the  $t$ -ratios were corrected for the difference in variance (1).

On the basis of the analysis for preventable accidents per 5,000 miles between the accident and non-accident groups, those six variables having the most significant  $t$ -ratio were chosen to be analyzed by the Wherry-Doolittle Shrunken Multiple Correlation Technique (3). This type of analysis gives the maximum coefficient of correlation between the test variables and the criterion score after a correction has been made for the chance error added by each variable. The first step in this technique is the computation of inter-correlations among the variables included and the criterion (in this case, number of preventable accidents per 5,000 miles).

### Results

Tables 1 and 2 show the condensed results of the study (only those variables showing significance out of the total of 43 included

Table 1  
Variables Showing Significant Differences Between Upper Half, Accident Group  
and Lower Half, Accident Group

Variable	$t$	Difference in favor of
Preventable accidents		
Systolic Blood Pressure	2.397*	Lower half
Far Acuity, Right Eye	2.448*	Lower half
Far Acuity, Left Eye	2.231*	Lower half
Near Acuity, Left Eye	2.411*	Lower half
Kuder Mechanical Interest	2.268*	Lower half
MMPI Hypomania	2.226*	Lower half
Far Vertical Phoria	2.802**	Lower half
Near Acuity, Both Eyes	2.746**	Lower half
Kuder Artistic Interest	10.332***	Upper half
Kuder Literary Interest	5.181***	Upper half
Non-preventable accidents		
Marital Status	2.690*	Lower half
Far Acuity, Both Eyes	2.526*	Lower half
Near Acuity, Both Eyes	2.270*	Lower half
Near Acuity, Right Eye	2.456*	Lower half

\* =  $P < 0.05$ .

\*\* =  $P < 0.01$ .

\*\*\* =  $P < 0.001$ .

Table 2  
Variables Showing Significant Differences Between Accident Group and Non-accident Group

Variable	<i>t</i>	Difference in favor of
Preventable accidents		
MMPI Hypochondriasis	2.011*	Non-accident group
MMPI Masculinity	2.136*	Non-accident group
Kuder Literary Interest	2.358*	Accident group
MMPI Psychasthenia	2.649**	Non-accident group
Kuder Artistic Interest	3.404***	Accident group
Non-preventable accidents		
Marital Status	2.413*	Non-accident group
Depth Perception	3.040**	Non-accident group

\* =  $P < 0.05$ .\*\* =  $P < 0.01$ .\*\*\* =  $P < 0.001$ .

in the study are mentioned for the sake of brevity). In the tables, "difference in favor of" refers to which half of the sample distribution has the higher mean score. With respect to Marital Status and Systolic Blood Pressure, higher score means that the driver is married and has higher systolic blood pressure than the mean for the sample.

As a result of the Wherry-Doolittle Shrunk Multiple Correlation Analysis, it was found that only four of the six variables could be used to obtain the maximum multiple coefficient of correlation; these four were: Kuder Literary Interest, MMPI Hypochondriasis, Masculinity, and Schizophrenia. The resulting multiple coefficient of correlation was 0.36. This coefficient is statistically significant at the 1 per cent level of confidence. The resulting prediction equation is:

$$X_c = 0.00234834X_{35} - 0.00376978X_{42} \\ - 0.00183060X_{39} + 0.00271910X_{31} \\ + 0.14349491.$$

In this equation the symbols have the following meaning:

$X_c$  = the predicted criterion score, number of preventable accidents per 5,000 miles

$X_{35}$  = the MMPI Hypochondriasis score

$X_{42}$  = the MMPI Schizophrenia score

$X_{39}$  = the MMPI Masculinity score

$X_{31}$  = the Kuder Literary Interest score

The standard error of a predicted criterion score is 0.1007.

### Discussion

The analysis indicates that the following variables—Systolic Blood Pressure; Far Acuity, Right Eye; Far Acuity, Left Eye; Kuder Mechanical Interest; MMPI Hypomania; Far Vertical Phoria; Near Acuity, Both Eyes; Kuder Artistic Interest; and Kuder Literary Interest—are related to preventable accidents. That is, those drivers having high scores on all these variables, with the exception of Kuder Artistic and Literary Interests, tend to have a lower number of preventable accidents. Those drivers having high scores in Kuder Artistic and Literary Interests tend to have a higher number of preventable accidents.

In looking at the analysis of the non-preventable accident data, we see that the psychological test data variables have dropped out, and that only the following appear to be related to non-preventable accidents—Marital Status; Far Acuity, Both Eyes; Near Acuity, Both Eyes; Near Acuity, Right Eye; and Depth Perception. These all bear the same relationship to the non-preventable accident data; that is, the higher the score, the fewer the number of non-preventable accidents.

It might be possible to postulate from the results of this study that one of the main distinctions between preventable and non-pre-

ventable accidents is the fact that preventable accidents seem to be related to psychological test data. Similarly, it might be said that such sensory capacities as the visual skills and certain personal history data such as Marital Status seem to be related to non-preventable accidents.

On the basis of the Shrunken Multiple Correlation Analysis, it was shown that by using the four variables—Kuder Literary Interest, MMPI Hypochondriasis, Masculinity, and Schizophrenia—a reasonably satisfactory prediction of the number of preventable accidents per 5,000 miles could be made. This part of the study seems to be quite significant. It would suggest that the psychological test variables, particularly those dealing with personality traits, are the variables which contribute to the discovery of those drivers having a high preventable accident rate. It might be said that this would seem to bear out the results of the first part of the study.

It should be mentioned that considerable selection has already taken place before these particular drivers are hired by the company. The mere fact that they attended the North Carolina State College Driver Training School makes them a select or special population, and therefore unlike what might be called representative of the general population of commercial truck drivers. We do not have any "extreme" cases in the sample, and it is these who seem to contribute a great deal to any analysis, especially one in which an attempt is being made to isolate variables which are contributing to job success.

It must also be noted that the apparent significance of the *t*-ratios reported is inflated. This is due to the fact that those reported are selected from a total of 43 for each phase of the analysis of significance of differences.

### Summary and Conclusions

This study has been concerned with studying the relationship of certain psychological and personal history data to the accident rec-

ords of a sample of commercial truck drivers. The accident data used were of two types—preventable accidents and non-preventable accidents, and were equated for the number of miles driven.

The subjects for the study were 104 commercial truck drivers employed by a large East Coast trucking concern.

For each of the criterion groups two types of analyses were made: (a) the significance of differences between the upper and lower halves of the accident group was computed; and (b) the significance of differences between the accident and non-accident group was computed. In addition, a Wherry-Doolittle Shrunken Multiple Coefficient of Correlation was computed for four of the variables and the criterion score based on the first analysis of significance of differences.

From this study the following conclusions can be drawn:

1. A difference seems to exist between preventable and non-preventable accidents.
2. Psychological traits, as well as sensory capacities, are important in analyzing the accident liability for preventable accidents, while only personal history data and sensory capacities seem to be important in analyzing the accident liability for non-preventable accidents.

This study also demonstrated the applicability of a technique for controlling exposure to accident hazard; that is, using the number of accidents per unit mileage rather than just the total number of accidents.

Received January 30, 1953.

Early publication.

### References

1. Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
2. Guilford, J. P. *Fundamental statistics in psychology and education*. (2d Ed.) New York: McGraw-Hill, 1950.
3. Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Company, 1940.

## Applied Psychology in Action

### A New Management Tool for Top Executives

The first issue of *Social Science Reporter* appeared on April 15, 1953, and is to be issued semi-monthly. It is an effort to build a bridge of better understanding between social scientists and the business leaders of America. The editor and publisher is Rex F. Harlow, 365 Guinda Street, Palo Alto, California, who wishes to be kept informed on the research that is being done. His staff will also seek to secure material for publication through personal interviews with scientists, the reading of scientific periodicals and attendance at scientific meetings.

The first issue contained abstracts of ten research projects such as: management organization and motivation (W. F. Whyte); management and the individual in an organization (E. Wight Bakke); multi-relational sociometric survey (I. R. Wechsler, R. Tannebaum, and E. Talbot); administering changes (Harriet O. Ronker and P. R. Lawrence); executive retirement (H. R. Hall); management-employee communication (Nat. Soc. of Prof. Engineers); improving supervisors (N. F. Maier); pension-plan policies and practices (M. Puchek); changing attitudes (L. Festinger and H. Kelley); mass persuasion (D. Cartwright).

### Background of an Industrial Psychologist

Excerpts from a letter and enclosure of personal data may be of interest and value to all concerned with putting psychology to work in business and industry. After congratulating Dr. Marion A. Bills for her paper on "Our Expanding Responsibilities" and for developing an *Applied Psychology in Action* section in the April, 1953 issue of the *Journal of Applied Psychology*, Dr. John F. Michael wrote: "More than ten years ago I decided to become a psychologist with both feet solidly entrenched in industry. Consequently I have had a series of jobs each selected with the criterion being the extent to which it would

contribute to a comprehensive industrial and business background. The two page enclosure, which might interest you, described this experience in greater detail.

"This attitude toward my professional relation to industry found great support in L. L. McQuitty's article in *American Psychologist* of several years ago. He also felt the great need of first being in demand to business not solely because of one's psychological background but mainly because of one's business background. The implications for the problem of communications are obvious. This is notwithstanding the fact that one of our largest consulting firms in industrial psychology in selecting their own staff maintains they do not particularly consider any possible business experience of the psychologists they hire.

"It seems fashionable to decry the small published output from psychologists spending all their time in industry. I fail to agree with the typical reasons offered by others for this. The main cause appears to me to be not any dire result of greater income, or any decrease in the desire to contribute to the general advance of the field, but merely the result of the lack of time and the need to sell to management the 'problem' of research. As a result it is almost impossible to do research as defined by other colleagues who have more time. An interesting support of this time factor might be the reading habits of the psychologist in industry. My own observation with reference to myself and others is that there is an inevitable decrease in professional reading from the dropping of the journal club subscriptions to the selection of a few journals none of which are completely read. I feel that such a decrease has not taken place so widely among our academic colleagues.

"The proposed feature is a most excellent suggestion. Comments to be accepted for publication should not be restricted too much with reference to content. Observations worthy of possible restatement into testable hypotheses, needs for specific techniques, notes as to application of psychological principles to unique problems in business are ex-

amples of possible contributions. In general, the basic characteristic should be brevity. This proposed section should give the psychologist in industry a voice by which to record some of his many varied actions which might be of interest and value to others. The thought of writing a comprehensive article likely has stifled the expression of many a good idea which, perhaps, would have required only several short sentences anyway. These ideas should not be lost.

"I fully realize nothing has been said to add any value or new points to Dr. Bills' speech. It is just a case of my being very pleased with her excellent discussion of the psychologist in industry with special reference to those of us who do not wear a tag which says 'psychologist.'"

Here are the high-lights of this industrial psychologist's background and training:

*Age:* 35; *Education:* AB, BS in LS, MA, and BBA, Western Reserve University; PhD, Ohio State University in 1952. *Majors:* Bibliographic research, general business administration, psychological aspects of marketing and merchandising.

*Work History:* (1) Business reference assistant, city public library, 3 years; (2) Research analyst, procurement division, Quartermaster Depot, 3 years; (3) Market analyst, for a large tire and rubber company, 3 years; (4) Jobs in The Lincoln Electric Company of Cleveland, Ohio, since January 1951:

automatic saw operator, plant layout detailer, assistant purchasing agent, factory engineer, management analyst, and at present management engineer.

*Professional Organizations:* (1) American Psychological Association, Associate member since 1947; (2) American Marketing Association; (3) Cleveland Psychological Association; (4) Personnel Psychologists of Northern Ohio.

### Noise and Absenteeism

A study reported by Dr. de Almeida is concerned with the maintenance of a work force of tabulating machine operators. He was called in because of absenteeism in a tabulating department of 125 workers. Noise was found to be a real source of trouble. He found that this could lead to aural lesions and neuropathic troubles. He recommended a rearrangement of equipment, placing machines on rubber supports, and "Cellotex" insulation of walls. These changes dropped the decibel average from 42 to 14. He states that 50 to 60 decibels, continued, causes permanent damage to the ear. Absences were reduced by 23 per cent, or 430 man/days in 6 months.—H. Ribeiro de Almeida, M.D. (Sao Paulo, Brazil). Influence of electric punch card machines on the human ear. *Archives of Otolaryngology*, 1950, 51, 215-222.

## Book Reviews

Heneman, H. G., Jr. and Turnbull, J. G. (Editors). *Personnel administration and labor relations: a book of readings*. New York: Prentice-Hall, Inc., 1952. Pp. xiv + 434. \$3.95.

Pigors, P. and Myers, C. A. (Editors). *Readings in personnel administration*. New York: McGraw-Hill Book Company, Inc., 1952. Pp. xii + 483. \$4.50.

Although not explicitly described as such, these two publications serve as supplementary reading sources for Yoder's *Personnel management and industrial relations* and Pigors and Myers' *Personnel administration*, respectively. They reflect the same similarities and differences as the basic texts themselves. Both attempt to present selected readings dealing with the various aspects of the field, the former stressing the "practical operating problems that confront the personnel man, labor relations man, or union leader in everyday life" and the latter "the philosophy of personnel administration, its basic problems and limitations, as well as criticisms and doubts raised by union leaders."

Heneman and Turnbull group their selections under four parts: The Setting of Industrial Relations; Personnel Administration; Labor Relations; and Research and Evaluation. They present 170 readings in all, many of them not being the entire article but excerpts selected to deal specifically with the topic under consideration. This method of presenting readings has certain advantages and disadvantages. It reduces overlap and duplication of content but at the same time requires a clear over-all structure to tie the excerpts together. This the authors have attempted to provide by including introductory statements before each part and by presenting a brief abstract, at the beginning of each chapter, of the main points of each reading and its contribution to the topic being covered. Despite the editors' care, some of the excerpts still "hang in mid-air" and the reader may feel a need for the omitted material which structured the original articles.

In contrast, Pigors and Myers present only 46 readings, organized under six parts: Nature and Scope of Personnel Administration; Analyzing and Handling Personnel Problems;

The Foreman; Building and Maintaining Work Teams; Wage and Work Assignments; and Employee Services and Programs. The individual readings tend to be longer and to be more self-contained (than those in Heneman and Turnbull) and appear to be the complete original rather than excerpts. They also, however, are organized under sub-topics and each part includes an introductory statement, by the editors, which interprets the reading and shows its contribution to the topic being treated.

It would be impossible to rate one of these publications as better than the other; they are just different. Although both provide adequate coverage of the general field and have eight authors in common, there is no selection which appears in both. Each has its favorite sources—Heneman and Turnbull the University of Minnesota Industrial Relations Center, and Pigors and Myers the Harvard Business Review—but both draw upon American Management Association publications, the Personnel Journal, recent books, and related fields.

Both of these publications are worth being in the industrial psychologist's library. They have a double value—for what is in them and for what is not. The readings themselves are drawn from a wide variety of sources and disciplines and indicate clearly that the industrial psychologist, to be truly effective and to be able to communicate with management, must become conversant with a wealth of literature outside of psychological journals. That psychologists are becoming recognized as contributors to the field is indicated by the fact that slightly over 10% of the authors in each book are psychologists. However, these are the psychologists who have written articles for the personnel journals or who have participated in conferences run by personnel management groups. It seems clear that psychologists cannot be too modest and coyly wait for management to wade through psychological journals to find out what psychology has to offer. At any rate, only a very few of the readings in these two publications are taken from psychological journals.

It is also instructive to note in which areas authors from psychology are used. As one

might expect, Selection and Placement, Training, and Research are those in which the psychologist is recognized as making the most unique contribution, but other areas are beginning to be influenced, areas such as morale, safety, job analysis, and even administration. However, if these two publications are a guide, psychologists have still not made themselves felt as contributors to an understanding of labor-management relation, incentive plans, union activities, grievances, discipline, etc. Either we have nothing useful to offer on these problems or have not made it readily available to nonpsychologists. I hope it is the latter.

Albert S. Thompson

Teachers College,  
Columbia University

Walker, C. R. and Guest, R. H. *The Man on the assembly line*. Cambridge, Massachusetts: Harvard University Press, 1952. Pp. 175. \$3.25.

The authors have contributed this pilot study in an effort to increase our general knowledge of the adjustments made and the satisfactions derived by workers on an assembly line.

It is a thorough, well-organized inquiry into the expressed feelings and attitudes of these men regarding many facets of their work. The book describes in detail the work climate at Plant X which is an automobile assembly plant. The method of investigation was both qualitative and quantitative, giving a clear picture of attitudes toward seven characteristics of the work. These, as listed by the authors, were: "the worker's immediate job, his relations to fellow workers, pay and security, his relation to supervision, general working conditions in the plant, promotion and transfer, and his relation to the union." These attitudes were determined by interviews, questionnaires, and the study of overt behavior.

The results of this investigation are not new or startling. It is a typical study of attitudes of employees toward their work situation, using the usual attitude research methodology showing us which aspects of the work were most liked and disliked.

Unfortunately, the authors do not emphasize the main reason the employees *liked* their work. "Pay." This criticism can be levelled

at many other social scientists of course. Perhaps it is because this aspect is considered out of their realm and they can't do much about it. On the other hand, they have come up with some realistic suggestions for alleviating, at least in part, the main *dislikes* which these men expressed concerning their work which was the immediate job content. This further inquiry into the specific aspects of the job content itself which the men objected to is the only unique contribution of the book. Evidently, it was not hard work itself but the mechanical pacing and the repetitiveness of the job to which they objected most strenuously. Walker and Guest have proposed what appear to be practical suggestions for minimizing these objections.

The essence of the problem seems to be that we have progressed too rapidly in the technique of mass production and it is time now to stop and re-organize our thinking to take into consideration the worker himself in order to solve some of the problems created. Many industrial engineers and members of plant management, however, are going to disagree until we can prove its economic effectiveness.

John M. Cook

Radio Corporation of America,  
RCA Victor Division, Camden, N. J.

Laird, D. A. and Laird, Eleanor C. *Practical sales psychology*. New York: McGraw-Hill, 1952. Pp. xii, 291. \$4.00.

The formula adopted by the authors of this book is clear and easy to follow up to a point. Take self-administering tests of twelve to eighteen items each and label them self-sufficiency, dominance, self-confidence, social mixing, friendliness, hygienic habits, fatigue, tactfulness, kindness, sympathy, warm-heartedness, optimism, ability to meet emergencies, considerateness, snobbishness, egotism, fault-finding, self-centeredness, and hot-temperedness. This will show the reader "where you stand in some qualities which are important for the salesperson" (p. 117). It will also help to convince him that scientific psychology underlies your approach.

Sprinkle with authoritative statements such as: "About one person out of four has some characteristics which make him disagreeable" (p. 168). "Records show that graduates who received poor grades in college become as good

salesmen as those who had high grades" (p. 12). "Psychological studies have shown that the love of selling or 'sales drive' is essential for sales success" (p. 13). "Ninety-eight per cent of experienced sales people like selling better than any job" (p. 15).

Prove your points with "studies." For example, if you want to demonstrate that love of selling can be developed, quote a study which shows that the longer men have been selling fire protection equipment the better they like it. Do not mention the possibility that self-selection rather than development accounts for the results.

Lard with references to big names and stories of the Horatio Alger type. Mention N. W. Ayer, Marshall Field, John Wanamaker, Owen D. Young, Abraham Lincoln and Colonel Edward M. House. Tell how they succeeded. In the case of Colonel House quote him as saying, "Let the other fellow make the mistake first." Point out that "House gained political power in the Woodrow Wilson administration by following that policy" (p. 264).

Do not allow difficulties to disturb you. If the concept of the salesman implies a generality which does not exist, if what is true of selling soap is not necessarily true of selling bonds, say so and then forget it. If the criterion of success in selling is so difficult to establish that it represents a major problem to psychologists in that field, don't even say so. Ignore it.

On the other hand, abash reviewers such as this one by writing with a style and clarity that far exceeds what your more conservative brethren produce. Demonstrate that you have absorbed the best of the common-sense stereotypes about the sales situation and that you have a rare gift for organization and emphasis. Beat all of the "Sales Power in Five Easy Lessons" boys at their own game. Convince even the psychologists that, after all, salesmen are probably better off to read your book than most of the literature that is now given them.

But do not say that "The first attempts to teach something about human nature in relation to the customer produced a lot of plausible-sounding nonsense. Magazines and books were filled with 'how to sell' advice which made psychologists, who knew human

nature, laugh aloud" (p. 38). *Who's laughing?* S. Rains Wallace, Jr.

*Life Insurance Agency Management Association, Hartford, Connecticut*

Barlow, F. *Mental prodigies*. New York: Philosophical Library, 1952. Pp. 256. \$4.75.

This book is described on its title page as "An enquiry into the faculties of arithmetical, chess and musical prodigies, famous memorizers, precocious children and the like, with numerous examples of 'lightning' calculators and mental magic." Actually most of it is devoted to lightning calculators and to examples of the mathematical tricks, stunts and problem solving used in performances of "mental magic." The author is himself something of a "mathemagician," having given performances using his own mnemonic devices and calculating short-cuts. His interest in such matters dates back to the nineteen-twenties or earlier and has led him to interview several arithmetical prodigies. For the most part, however, his knowledge of prodigies is second or third hand.

The first chapter, 58 pages in length, is an interesting account of nineteen outstanding arithmetical prodigies of the last two and a half centuries, together with brief notes on twenty-two others less well known. The large majority of both groups were born between 1700 and 1870, and only seven of them since 1900. For the outstanding cases the author gives where possible date of birth and of death, nationality, age when the special gift became evident, amount of education, examples of calculating feats performed, and an estimate of general ability. Here he draws heavily from Scripture's 1891 article on "Arithmetical prodigies" in the *American Journal of Psychology*, and from Mitchell's 1907 article on "Mathematical prodigies" in the same Journal. As most of the cases reported upon antedated scientific psychology, their histories, with a few notable exceptions, are sketchy and poorly documented. The author cites uncritically hear-say evidence, articles from old newspapers and popular magazines, and writers on psychic research. He also has a few systematic biases, including: (1) the usual bias of one who has dabbled in psychic research; (2) an admitted "instinc-

tive dislike of precocious children"; (3) a tendency to exaggerate the one-sidedness of prodigies and to underestimate their general ability. The third of these may be an outcome of the second. For example, the Belgian boy Verhaeghe is described as "an adolescent of seventeen with the mental age of a babe of two years." Yet this boy, when examined by a committee of mathematicians in 1946, gave the fourth power of 1,246 in 10 seconds, the sixth root of 24,137,585 in 25 seconds, and the square of 888,888,888,888,888 in 40 seconds!

The arithmetical prodigies whose nationalities are given distribute as follows by country: Britain 10, France and Italy 7 each, United States 5, Germany 2, and 1 each for India, Ceylon, Greece, Belgium, Spain, Switzerland, Egypt, and Mexico. Britain's lead is no doubt due to the fact that the author has covered his own country more thoroughly than other parts of the world. Of the United States' five, two were Negro slaves (one of them born in Africa), and two were from the little state of Vermont. The author's list includes two very famous scientists (Gauss and Ampère), a father-son pair (George P. Bidder, Sr. and Jr.), and six university graduates (both Bidders, Gauss, Ampère, the American Safford who became a professor of astronomy, and Whately, an Archbishop of Dublin). Several others gave evidence of superior general intelligence. At the opposite extreme three were considered mental defectives, though the basis for such classification is dubious in two of the three cases. Several described as dull were almost certainly well above average in general ability. Of the 41 arithmetical prodigies, four were sheep herders in childhood, two were blind, two had 12 fingers and 12 toes, and one was born without legs or arms. The relatively high incidence of sheep herding and of physical anomalies is hardly explicable in terms of chance, nor is the sex ratio of 39 males to 2 females.

The book would have been more acceptable to psychologists if it had included only the chapters on History and Data (58 pages), The Calculations Considered (15 pages), Development (6 pages), Famous Memorizers (19 pages), Mnemonics (9 pages) and Mental Magic (53 pages). The remaining hun-

dred pages are a hodgepodge of naive comment on such topics as heredity and instinct, mental imagery, chess and musical prodigies, precocity and genius, and the subconscious.

The book is entertainingly written. The authenticated feats of calculation and memory which the author has recounted are fascinating and astonishing. The chapters on mental magic will intrigue many readers by the revelations given about such things as naming the day of the week on which a dated event occurred or will occur; short-cut methods for a great variety of computations, including among others squaring, cubing, root extraction, translating months or years into seconds and miles into inches or barleycorns; and directions for carrying out such parlor stunts as "think of a number," "draw any card," "a loan and a present," "change for a shilling," and numerous others. By showing that many feats which appear so difficult can be mastered by almost anyone with a moderately high IQ and a fair memory, these chapters can be expected to increase the number of amateur mathemagicians whose performances will probably amuse, mystify, and bore us in about equal proportions.

Lewis M. Terman

Stanford University

Dunsmoor, C. and Davis, O. *How to choose that college*. Boston: Bellman Publishing Co., 1951. Pp. 51. \$90 (cloth bound).

The rapidly expanding number of students considering college training makes the need for materials to assist them with their choices increasingly urgent. This book was written for high school students and their parents to discuss many of the questions which they typically have. It is simply written and attractively illustrated. The authors systematically consider, among other things, such problems as college requirements, application, financial plans, and making good once admitted. This is quite an order for fifty-one pages.

To cover this ground, the authors rely heavily on lists of things to do and avoid, on typical recommended high school curricula, considerable oversimplification, and frequent suggestions to see a counselor. The book would, therefore, seem to be best suited for

use in a guidance class; without expert fill-in, most students and their parents would probably either be left with unanswered questions or be misled.

Counseling plays an interesting dual role in this book. On the one hand, the student is left in rather large measure to decide for himself whether he has the right goals, interests, and abilities. This is in spite of the ample available evidence regarding the tendency to appropriate to the self desirable characteristics in such circumstances. On the other hand, the student is sent to the counselor for such things as assistance with writing of letters of application and, immediately on receipt of acceptances, for "review" of these. This suggests a dependency likely to plague college officials and counselors later on.

Properly handled, this book should make a useful contribution to the literature on guidance. But, because it has been almost oversimplified and condensed, it is doubtful whether it should be turned over to students as the sole and sufficient guide through a difficult terrain.

John W. Gustad

University of Maryland

Argyris, C. *An introduction to field theory and interaction theory*. New Haven: Yale Labor and Management Center, 1952. Pp. 71. \$1.00.

This brief 71-page pamphlet was originally prepared for members of the Labor and Management Center at Yale University. Presented in each case are some basic foundations of each theory, some assumptions underlying each theory, some of the goals of each theory, some basic operations suggested by each theory, and some of the more important concepts utilized in each theory. The writing is clear in so far as the somewhat turgid language of the authors, whose work is summarized, makes that possible, and simple in so far as the extreme complexity of the theories discussed allows. As a presentation of field theory and interaction theory, therefore, this book may be recommended.

There are several points, however, on which the critical reader may well feel uneasy. In the first place, the author makes no attempt to link up the two sets of theories he is dealing with. Is field theory compatible with

interaction theory? Do they deal with the same aspects of human behavior? Are there alternative descriptions of similar phenomena? Is there any evidence of an experimental kind to make a choice between them possible? No information is given on these vitally important points.

Again, in what way do the concepts and ideas of Lewin, on the one hand, and of Chapple, Arensberg, Whyte, and Homans, on the other hand, constitute a theory? In science the term theory is ordinarily used to denote a system of hypothetical constructs or intervening variables which enables us: (a) to summarize existing knowledge; and (b) to make deductions leading to testable predictions. This clearly is not the purpose of either field or interaction theory. Both are essentially semantic attempts, occasionally leavened by the use of unorthodox mathematics, to persuade the reader to accept unproven views and to make use of complex ways of stating what often appear to be obvious, commonsense notions. Argyris never comes to grips with the problem of proof and disproof of theories of this kind, nor does he attempt to show that they have a function to perform which could not equally well be performed in other ways.

This failure to provide any form of critical discussion is the major weakness of the book. The unwary reader might think from Argyris' presentation that Lewin's use of topological concepts, or hodological space, is acceptable to orthodox mathematicians. This, as far as I know, is not so, and one might have expected at least a brief reference to criticisms of these, as of many other points. A book of this kind will not be very useful to the initiated, who will be familiar with its contents in any case; nor will it help the uninitiated to obtain a balanced view of the theories proposed. The habit of stating a position without answering important criticisms of that position, or even mentioning that such criticisms exist, is unfortunately widespread in psychology. This book is an outstanding example of what, to the reviewer, appears to be a very bad practice indeed.

H. J. Eysenck

Psychology Department, Maudsley Hospital  
London, England

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Introduction to the Rorschach technique.* Robert M. Allen. New York: International Universities Press, Inc., 1953. Pp. 126. \$3.00.
- Introduction to exceptional children.* Revised edition. Harry J. Baker. New York: Macmillan, 1953. Pp. 500. \$5.00.
- Social psychology.* Hubert Bonner. New York: American Book Co., 1953. Pp. 439. \$4.25.
- Progress in clinical psychology.* Vol. I, Sec. 1. Daniel Brower and Lawrence E. Abt, Editors. New York: Grune and Stratton, 1952. Pp. 328. \$5.75.
- The individual and world society.* P. E. Corbett. Princeton: Center for Research on World Political Institutions, Princeton University, 1953. Pp. 59. Gratis.
- The MMPI: a review.* William C. Cottle. Lawrence, Kansas: School of Education, Univer. of Kansas Publ., 1953. Pp. 82.
- Logic and language.* Second series. A. G. N. Flew, Editor. New York: Philosophical Library, 1953. Pp. 242. \$4.75.
- The Grassi Block Substitution Test for measuring organic brain pathology.* Joseph R. Grassi. Springfield: Charles C Thomas, Publisher, 1953. Pp. 75. \$3.00.
- Initiating and administering guidance services.* S. A. Hamrin. Bloomington, Illinois: McKnight and McKnight, 1953. Pp. 220. \$3.00.
- Introduction to statistical methods.* Palmer O. Johnson and Robert W. B. Jackson. New York: Prentice-Hall, 1953. Pp. 394.
- Heredity in health and mental disorder.* Franz Josef Kallmann. New York: W. W. Norton Co., Inc., 1953. \$5.00.
- Problem drinkers can be helped.* G. N. Lansdown. Devon, England: Arthur H. Stockwell Limited, 1953. Pp. 72.
- Psychology of industrial relations.* C. H. Lawshe. New York: McGraw-Hill Book Co., Inc., 1953. Pp. 350. \$5.50.
- Comparative conditioned neuroses.* Roy Waldo Miner, Editor. New York: New York Acad. of Sci., 1953. Pp. 379. \$3.50.
- The natural superiority of women.* Ashley Montagu. New York: The Macmillan Co., 1953. Pp. 205. \$3.50.
- God, labor and management.* Alfred Morgan. New York: The William-Frederick Press, 1953. Pp. 28. \$1.00.
- Group psychotherapy.* Florence B. Powdermaker and Jerome D. Frank. Cambridge: Harvard Univer. Press, 1953. Pp. 615. \$6.50.
- The conception of disease.* Walther Riese. New York: Philosophical Library, 1953. Pp. 120. \$3.75.
- The universe of meaning.* Samuel Reiss. New York: Philosophical Library, 1953. Pp. 227. \$3.75.
- Philosophy and the ideological conflict.* Charles S. Seely. New York: Philosophical Library, 1953. Pp. 319. \$5.00.
- The retention of meaningful material.* Joseph Francis Sharpe. Washington, D. C.: Catholic University of America Press, 1952. Pp. 66. \$1.00.
- Capitalism overhauled.* Job Socius. New York: The William-Frederick Press, 1952. Pp. 79. \$2.00.
- Introduction to testing and the use of test results in public schools.* Arthur E. Traxler, Robert Jacobs, Margaret Selover and Agatha Townsend. New York: Harper and Brothers, 1953. Pp. 113. \$2.50.
- Retirement and the industrial worker.* Jacob Tuckman and Irving Lorge. New York: Bureau of Publications, Teachers College, Columbia Univer., 1953. Pp. 105. \$2.75.
- The science of color.* Committee on Colorimetry of the Optical Society of America. New York: Thomas Y. Crowell Co., 1953. Pp. 385. \$7.00.
- Evaluating research and development.* Irving R. Weschler and Paula Brown, Editors. Los Angeles: Institute of Industrial Relations, University of California, 1953. Pp. 104. \$1.65.
- Research Report of U. S. Naval School of Aviation Medicine.* Pensacola: U. S. Naval Air Station, 1953.

# Journal of Applied Psychology

VOL. 37, No. 5

OCTOBER, 1953

## The Weather and Other Factors Influencing Employee Punctuality \*

Roland E. Mueser

*The Pennsylvania State College*

On a beautiful warm unseasonal day in the middle of February 1951, the majority of usually sleepy students arrived bright and early for an 8:00 o'clock college class. Such promptness at this hour was as unusual as the spring day in February. The coincidence invited the comparison of weather and attendance. Did the early morning brightness stimulate these otherwise uninspired students? The hypothesis suggested itself: Increased light intensity might be causing early awakening, or it might hasten the morning routine of washing, dressing, and breakfasting. This study was undertaken to determine the correlation between early morning illumination and one indicator of human activity. Promptness in reporting to work was used as a criterion which might be accurately measured on a statistically significant population.

The personnel of an engineering research laboratory on the campus was chosen because attendance figures could be readily obtained. Only those employees who were scheduled to start work at 8:00 a.m. were selected and a standardized list was used for holding a constant sample. It was, however, impossible to rigidly limit the sample to the identical group for many practical reasons. Part of the employees were necessarily absent on business, vacation, or because of illness during intervals in the recording period. Eliminating their records was a prohibitive statistical task and

would also result in a drastically reduced population size. Actually a total of 144 individuals were on the standardized list. Of these, an average of 132.8 or 92.2% were at work in the Laboratory during the test period. By extending the study over a number of months a gross averaging effect has been achieved and should tend to minimize chance errors.

### Procedure

Employees of the Laboratory were checked in by guards at the gate and the time recorded to the nearest five minute interval. An average of 101.3 men and 31.5 women were timed six days a week from February 23, 1951 through May 14, 1951, a total of 69 working days. Data were recorded independently for men and women to allow for later comparison. The majority of employees drive to work in private automobiles and the remainder all walk. No public conveyance is employed for transportation, hence patterns due to standardized bus or train schedules are avoided. Similarly chance pattern disruptions, as might be due to a commuter train arriving late, are not present. Individual chance factors cannot, of course, be avoided. However, since an incident such as a flat tire affects no more than a single car pool, no large error is introduced by a single transportation mishap. The average distance traveled to work was 3.8 miles for employees, with approximately 66% living in State College where the drive or walk to work is less than a mile. Although the only primary division of the population is sex, inherently this tends to produce strong secondary selectivity. The women of the Labora-

\* The author wishes to thank Dr. William M. Lepley for his advice and aid. In particular, it was Dr. Lepley's classroom observation which was responsible for undertaking this project. In addition, the cooperation of the members of the Ordnance Research Laboratory administration and Meteorology Department is sincerely appreciated.

tory are mainly secretaries, clerks, typists, and a few technicians. The female employees are, therefore, an exclusively non-supervisory group. The group of 101 male employees is composed 60% of research scientists and administrators. Most of the remainder are made up of machine shop and male technical employees such as draftsmen and scientific assistants.

The weather data were obtained from the Meteorological Department of the College. The most important information for this study, a measure of light intensity, was obtained from the department's Eppley pyrheliometer. Readings were taken from the record of light intensity at half-hour intervals from 6:00 through 8:00 a.m. and a total of these values was used as a measure of the light intensity for the early morning. Applying these figures to all employees does, of course, involve the assumption that the general atmospheric conditions are the same at all homes as at the college. The closeness of most residences to the Laboratory and the averaging effect of a large sample would be expected to reduce errors due to this assumption.

*Nine other meteorological variables were observed at 7:00 a.m. These were included in the study in order that they might be considered as secondary influences on punctuality behavior.*

There is a question as to how early and how late arrival times are to be tabulated to obtain an average figure which is a sensitive indicator of deviations which atmospheric conditions might be expected to introduce. Although there is no obvious error in including all employees who arrived early, extreme lateness would seem due, a disproportionate fraction of the time, to purely chance factors rather than the interplay of a subtle meteorological influence. The flat tire, sick child, morning shopping trip, inoperative alarm clock, or the morning return from a business trip all introduce delays which would overshadow the effects being sought. A subsequent study of employees arriving very late verified the fact that these variations occur randomly. It was decided, therefore, to study most intensively the average arrival time of employees arriv-

ing at work less than 22.5 minutes late. In order that this group be balanced a similar limit was placed on early arrivals so that employees arriving after 7:37.5 a.m. but before 8:22.5 a.m. were studied.<sup>1</sup> Eighty-six percent of all men and 97 percent of all women arrived between these times. The average arrival times were calculated independently for each sex.

Conjecture would lead one to believe that the very early arrivals—i.e. those coming before 7:37 a.m.—would be extremely sensitive to meteorological influence since their attendance is not so keenly forced by the conformity-producing 8:00 a.m. deadline. Furthermore, as a group they might be expected to exhibit greater individuality. In other words, the early birds are more nearly free to do as they please and so should react markedly to any atmospheric condition which tends to produce stimulated or sluggish behavior. Because of this, arrival times of these early employees were also studied as an independent group. Since no women regularly came to work before 7:37 a.m., this computation was only possible for men.

#### *Distribution of Arrival Times*

The distribution of arrival times for men and women was computed for the period March 9 to May 14. Figure 1 shows the distribution of average arrival times of 32 women and 101 men.

The distribution is similar to that obtained by F. H. Allport (1) called a J curve because of the decrement characteristic of arrival times before the 8:00 a.m. deadline. The drop-off after 8:00 o'clock is also in agreement with earlier data, being steeper and in the shape of a reversed J. The greater steepness on the right side is expected since individuals arriving late are exhibiting non-conforming behavior. The conjecture seems reasonable that if the time of arrival were stipulated as 8:00 a.m. but no social or economic pressure

<sup>1</sup> Since the raw data of the study were grouped in five minute intervals all class divisions are on the half minute and the statistics have been computed on this basis. However, to avoid the awkward dangling .5, intervals are quoted here to the minute and the additional fraction is to be understood. In this case 7:37 and 8:22.

forcing conformity was applied, the resulting distribution might be a Gaussian or normal curve. Conversely as the factors to produce conformity—i.e., to get to work on time—be-

come more compulsive, not only will the curve be displaced to the left (as Allport points out) but the skewness of the distribution should become accentuated. For the extreme op-



FIG. 1. Average arrival time, March 9-May 14, 1953.

posite of "free-will" attendance there would be a situation where the punishment for lateness was so severe that a tardy employee would stay absent rather than risk lateness. Such a situation is not as fantastic as it sounds, for it is close to the actual circumstance where being late in catching an important plane or train is as bad as not arriving at all.

In the distribution illustrated here there is a marked difference in the average behavior of men and women. Whereas an average of about 8% of the men arrive extremely early, before 7:37, only 0.1% of the women come during this period. Similarly, more than twice as many men come extremely late, after 8:22, as women. Practically no female employee arrived at work later than 9:20 a.m., yet a few men drifted in every morning as late as 10:45 a.m. Beyond this point it is a moot question whether it is tardiness or half day absenteeism which is occurring.

The tendency for the attendance characteristics of men to be more widely distributed than women is believed due to the high proportion of male research and professional workers. Research is traditionally an occupation catering to individualism and personal work habits. Even in a highly structured situation where it is generally expected that regular hours will be kept, some of the technical employees do not take attendance rules too literally. The tendency towards pronounced earliness would seem to be due to the greater personal job interest. This is an understandable manifestation among men where working is a career. Almost all the women employees are performing less stimulating service jobs and few expect to work longer than a few years.

The time for men and women coming between 7:37 and 8:22 was averaged for each working day in the test period. If the time which these employees arrive at work is being influenced by a common factor, the mean arrival times of the two sexes should have a positive correlation. Indeed, such a test bears a resemblance to split-half methods of computing test reliability. Actually the average arrival times of men and women have a correlation coefficient of .43 with a level of sig-

nificance of 0.02%. Although the two groups vary in a similar manner from day to day the mean arrival time for men, 7:55.52 a.m., is 2.6 minutes earlier than that for the women.

It would be expected that precipitation would tend to make employees tardy since the majority drive to work, roads become slippery under these circumstances, and visibility is impaired. However, a comparison of rainy mornings with dry ones failed to reveal any significant difference due to this factor (see Table 1).

Table 1

Effect of Rain on Mean Arrival Time of Employees Arriving Between 7:37 and 8:22 a.m.

	Fair Weather (25 days)	Precipitation (44 days)
101 Men	7:55.80	7:55.39
32 Women	7:58.17	7:58.06
True Average	7:56.36	7:56.02

Since it is difficult to imagine that on the average no impairment of driving conditions existed due to precipitation, it appears that employees foresightedly take bad driving conditions into account and tend to compensate for the circumstance. It is also possible that some phenomenon adjunctive to rain has a reverse effect and tends to produce early attendance. Later results lend credence to this hypothesis.

### Weekly Cycle of Punctuality

The popularity of the expression "blue Monday" and a general totaling of introspective reports following any weekend would lead one to believe that the day of the week might have an influence on work attendance figures. A plot of the mean arrival time for the employees as a function of day of the week is given in Figure 2. It can be seen that there is agreement between the two groups with the exception of a let-down among men on Wednesdays. If punctuality can be considered a criterion of general feeling tone it is apparent that Monday is indeed "blue," people are hitting their stride by midweek, and tend to be more tardy as the weekend approaches.

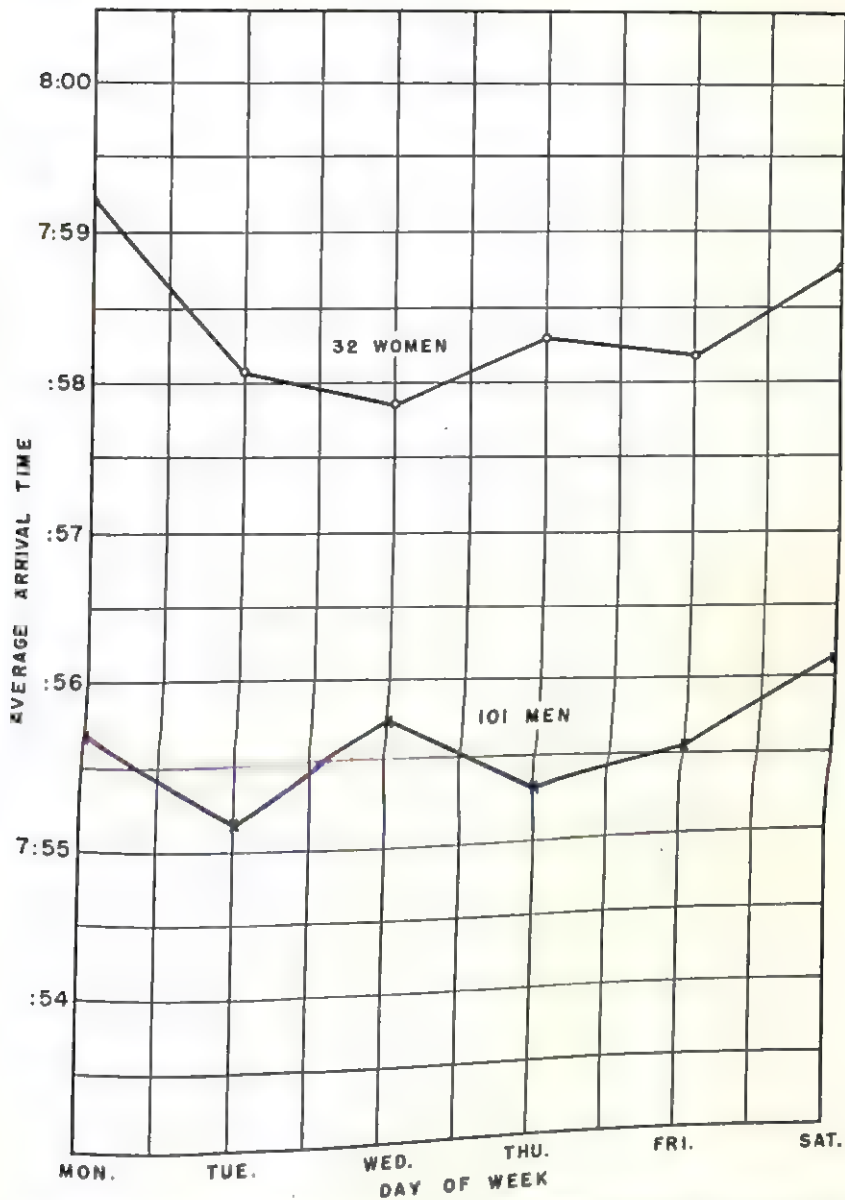


FIG. 2. Weekly cycle of average arrival time of people coming between 7:37 and 8.22 a.m.

A breakdown of all employees into time arrival groups per day illustrates a cross-pattern in punctuality habits as is shown in Figure 3. Fewer employees arrive *Just on Time* and more *Late* as the week progresses and people coming far ahead of time follow a different pattern from those arriving just a few minutes early. The curves are primarily of interest because they illustrate that the *shape* of the arrival distribution given in Figure 1 will vary somewhat as a function of the day of the week.

### Meteorological Effects

A figure representing early morning brightness was obtained by measuring the light intensity recordings of an Eppley pyrheliometer at half hour intervals from dawn until 8:00 a.m. The sum of these values "*I*" gives a rough integration of total morning brightness. Over the February to May period covered by the study the average value of *I* gradually increased. However, there was no significant change in employee arrival times indicating

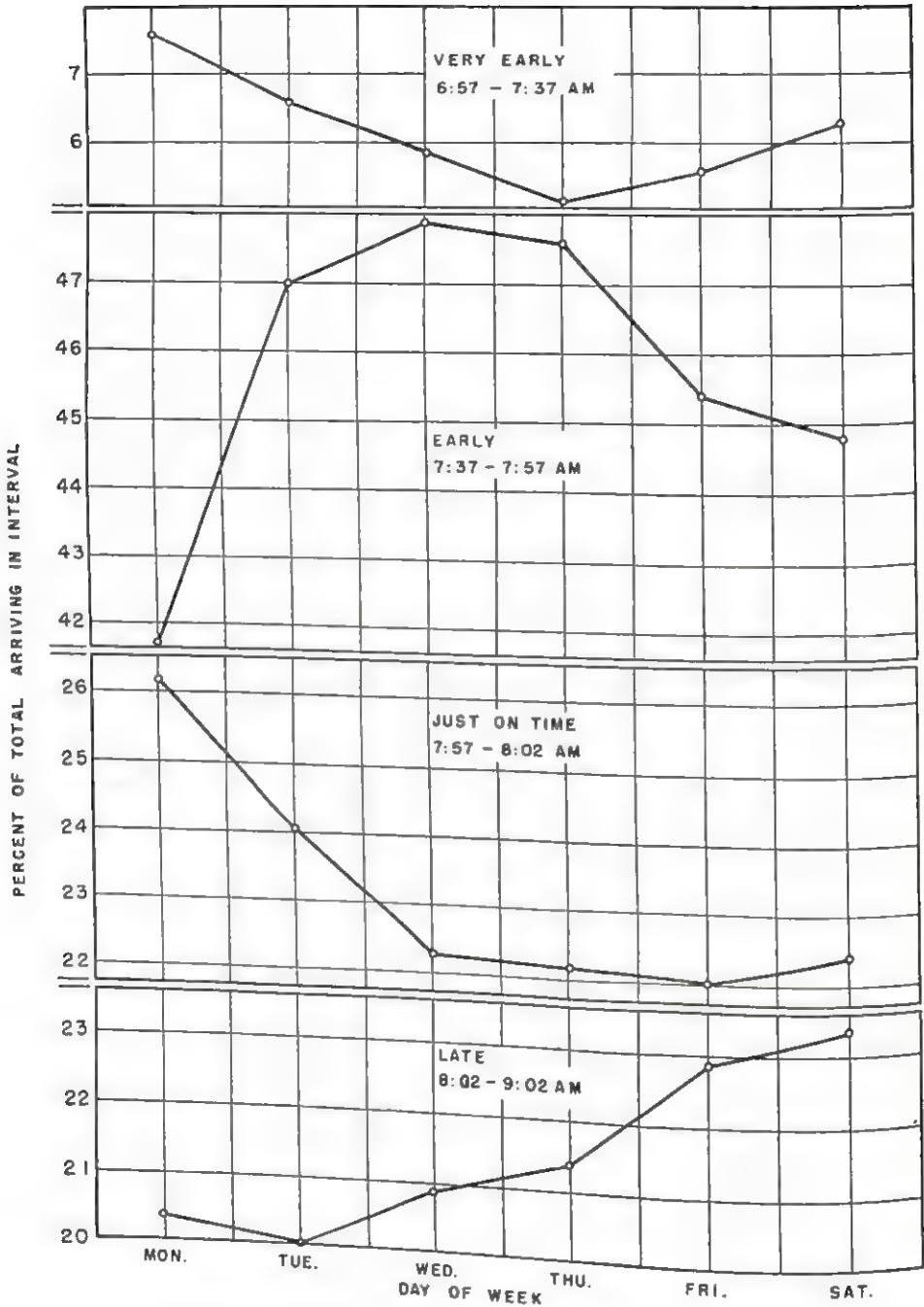


FIG. 3. Weekly cycle of punctuality for men and women.

adaptation to the seasonal light change. In general, however, the day-to-day fluctuations were far greater than the seasonal change, see Figure 4. Because of the wide range of values, and because psychophysical brightness discrimination is a relative rather than absolute

phenomenon, light values have been considered logarithmically (2). The correlation between the average daily arrival times and light intensity,  $10 \log I$ , is given in Table 2. In general the correlations between light

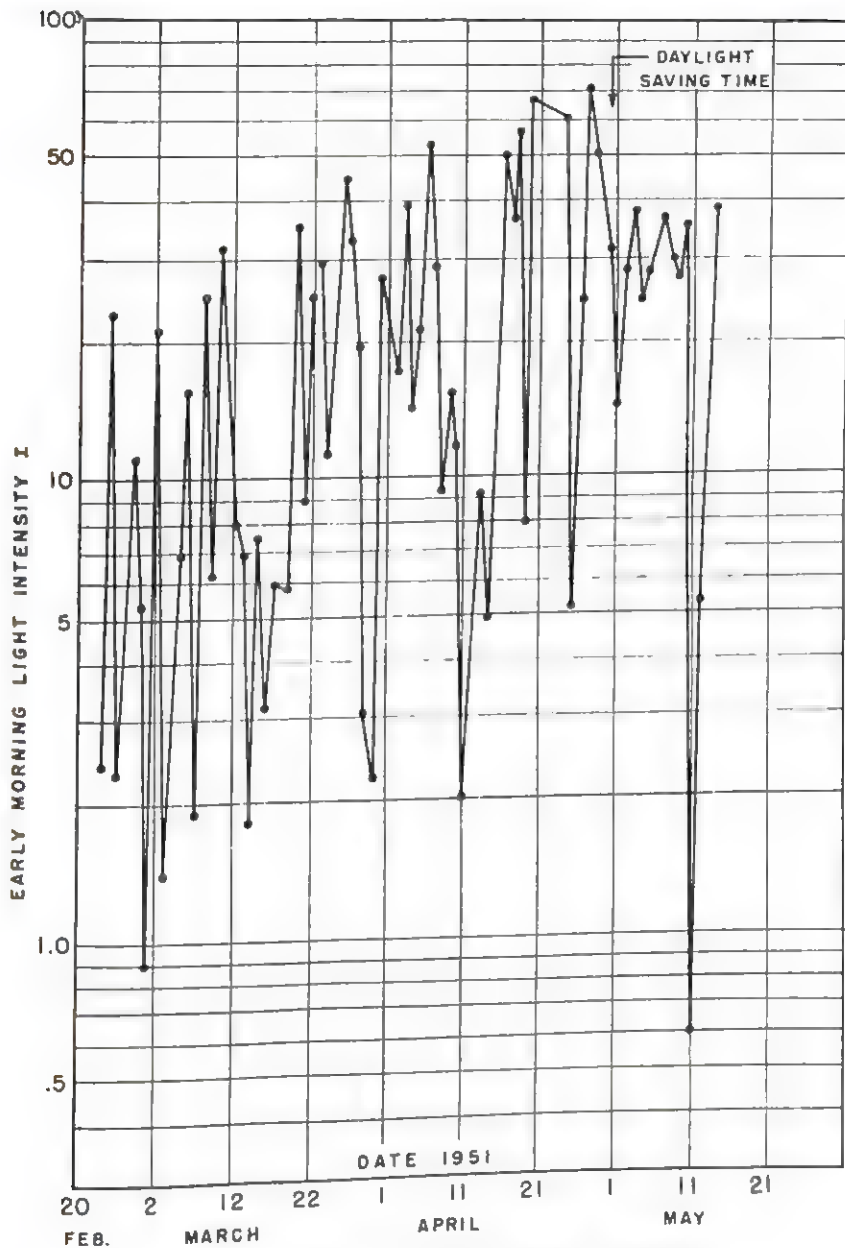


FIG. 4. Daily light intensity from dawn to 8:00 a.m.

and promptness are significant at about the 10% level but surprisingly in an inverse manner from that originally expected. Thus on the average both men and women arrive at work significantly earlier when the morning is dull and later when it is bright. The daily mean arrival times for men were grouped into thirds and those for women into fifths. From Figure 5 it is apparent that the reaction was similar in both men and women. The average

arrival time of women fluctuated over a much wider interval than that of the men. The men who arrived Very Early reacted in a contrary manner to the morning light stimuli just as they did with respect to the weekly cycle in Figure 3.

The data were also examined to determine whether any of the following meteorological conditions might be a factor influencing the average arrival times: (1) Corrected baro-

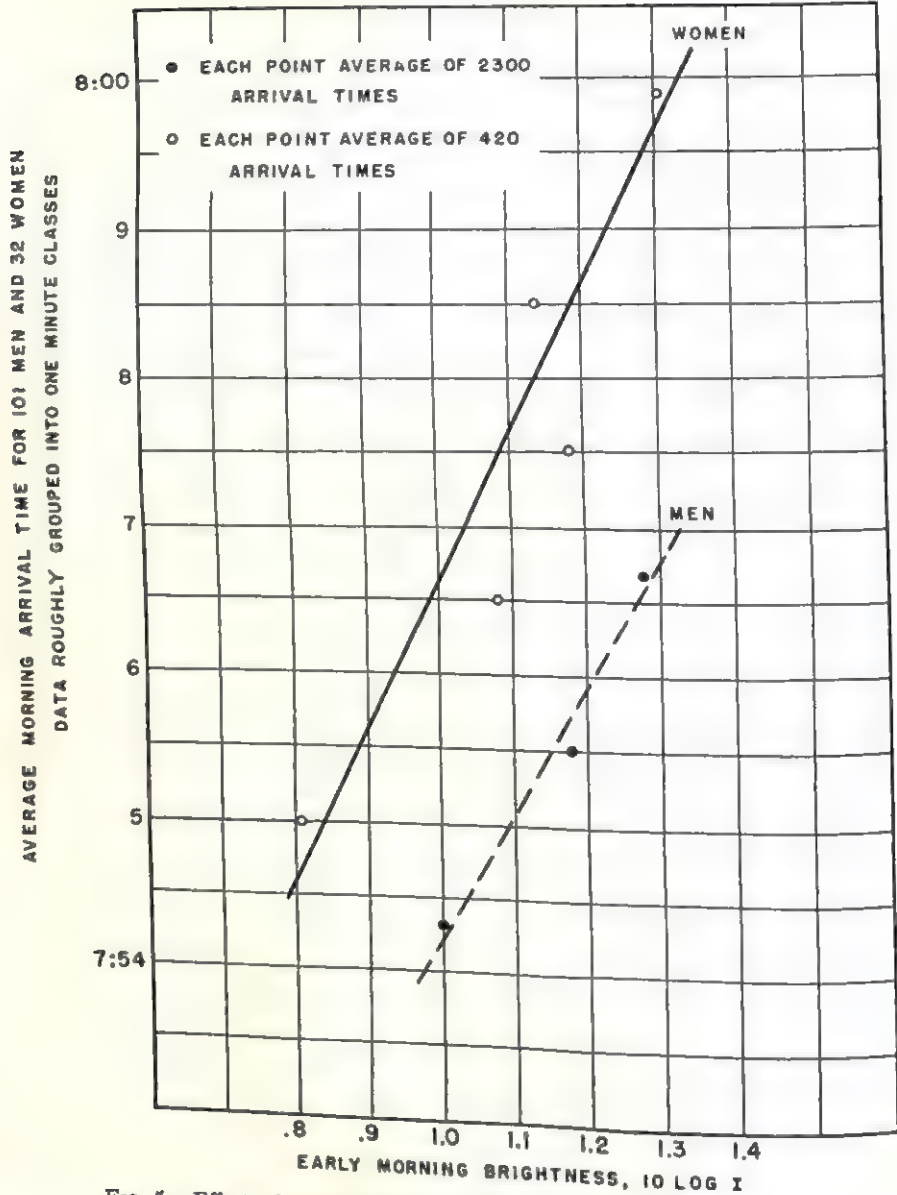


FIG. 5. Effect of light intensity on average time of arrival at work.

metric pressure at 7:00 a.m.; (2) Barometric tendency in last 20 hours with respect to direction and magnitude; (3) Amount of barometric fluctuation, i.e. atmospheric pressure roughness existing in the previous 20 hours (roughness was defined as pressure variations lasting less than an hour and including tendency reversals); and (4) Change in light intensity in the preceding 24 hours (both direction of change, and curve steepness were considered).

Mills (5), Winslow (6), and others have noted various psychological effects of barometric pressure on human beings. However, none of the listed factors showed a significant correlation with the punctuality criterion. The highest correlation coefficient obtained was +.12 between the barometric pressure and the average arrival time of women. A slight correlation in this direction would be expected as arising from the fact that dull overcast mornings are more common when the

Table 2  
Correlation of Employee Punctuality and  
Early Morning Light Intensity

	Morning Brightness, 10 log I	
	Correlation Coefficient	Level of Significance
Average arrival time most Men (7:37-8:22 a.m.)	+ .22	7%
Average arrival time most Men after correction for weekly cycle	+ .22	7%
Average arrival time most Women (7:37-8:22 a.m.)	+ .16	19%
Average arrival time most Women after correction for weekly cycle	+ .24	5%
Average arrival time very early Men (before 7:37)	- .26	3%

barometer is low. The correlation coefficient linking change in barometric pressure and punctuality was only .01 and other factors had similarly low values.

Numerous studies have shown that temperature and humidity factors are of physiological importance (4). Nevertheless, over the period studied, the environment of the subjects was almost entirely controlled by artificial means, home furnaces, car heaters, etc., rather than by the prevailing meteorological conditions. No important correlation of such factors as temperature, humidity, wind direction or wind velocity with employee attendance was evident. At this time of the year people are largely protected from direct weather influence by house walls and enclosed cars. These enclosures probably reduce all psychological and physiological weather manifestations. However, atmospheric pressure and light intensity would seem to penetrate such barriers to a greater extent than wind, cold, or humidity.

#### Discussion and Summary

The data studied here covering 8000 arrival times over a three-month period indicates that as a group, employees arrive at work in a pattern apparently inversely dependent on the brightness of the morning light. Such a

trend appears equally true for both sexes, with the exception of about 6%, all males, who come to work far ahead of the official starting time. Since their early arrival in itself sets these 6% as rather an individualistic group, it is not surprising to find that their reaction to both the weekly "fatigue" cycle and the light intensity is the converse of the rest of the workers.

In general, these reactions by employees would seem to reflect on their attitude about their jobs rather than serve as any reliable indication of feeling tone or satisfyingness (3). Thus it is easy to imagine that when it was sunny and beautiful outside the chore of earning a livelihood was put off. Perhaps a few fathers played a little longer with their children or paused to sniff a crocus on their way to the car. On the other hand on a dark, dismal morning, more often than not, the tired secretary and sleepy engineer drank their coffee more quickly and set off to work promptly and without fanfare.

Those men who arrive very early would seem to regard their work differently than the majority. One can only conjecture that they eagerly hurry to the job in the early morning sunshine, anxious to start another day of activity while their compatriots dawdle an extra two minutes admiring the blue skies. A world which produces both Stoics and Epicureans should not find such diverse behavior surprising.

Received November 13, 1952.

#### References

1. Allport, F. H. *The J-curve hypothesis of conforming behavior*. In *Readings in social psychology*. New York: Henry Holt, 1947.
2. Bartley, S. H. *Studying vision*. From *Methods of psychology*. New York: Wiley and Sons, 1948.
3. Bills, A. G. *Studying motor functions and efficiency*. From *Methods of psychology*. New York: Wiley and Sons, 1948.
4. Hirsh, J. *Comfort and disease in relation to climate. Climate and man*. Washington, D. C.: U. S. Government Printing Office, 1941.
5. Mills, C. A. *Medical climatology*. Springfield, Ill.: Charles C Thomas, 1939.
6. Winslow, C.-E. A. and Herrington, L. P. Subjective reactions of human beings to certain outdoor atmospheric conditions. *Heat., Pip-ing & Air Conditioning*, 1935, 7, 551-556.

## Prediction of Turnover Among Clerical Workers

Philip H. Kriedt and Marguerite S. Gadel

*The Prudential Insurance Company, Newark, N. J.*

Companies like The Prudential Insurance Company which hire a large number of High School girl graduates to do routine clerical work frequently have a turnover problem. We find that among the High School girls we hire each year some become permanent employees and make a career of their jobs. A larger number work for a few years and then quit to become housewives and raise a family. Both these groups we feel are good investments. There is a third group of new employees which concerns us, however. They are the girls who leave in a year or less to take other jobs or to go to college. These we consider to be a turnover problem.

We have done several investigations to see if we can reduce our turnover rate by determining at the time of employment whether or not a girl is a good turnover risk. Some of our most recent findings from this research are summarized in this article.

### Predictor Measures

All High School girls hired in June, 1951, were given an experimental battery of tests and questionnaires selected as possible predictors of turnover. The battery included a measure of intelligence, a measure of clerical aptitude, an interest questionnaire, a biographical data blank, and a job preference questionnaire.

1. General ability or intelligence was measured by two tests: Vocabulary and Arithmetic Reasoning. Scores for these two tests were combined for purposes of predicting turnover.

2. Clerical aptitude was measured by four tests: Name Checking, Number Checking, Dotting, and Letter-Digit Substitution. These four scores were combined to give a single clerical speed test score.

3. Interest scores were obtained from a questionnaire developed by the Company consisting of 285 items similar in form and content to those used by Strong. A key to predict turnover, consisting of 15 items scored by unit weights, was developed from data obtained in a previous study. A longer key consisting of 43 items did not cross-validate as well as the shorter key. The key identifies poor turnover risks as girls who like artistic, literary, scientific, selling, and social serv-

ice activities, and who dislike manual, mechanical and clerical activities.

4. Biographical information was obtained in a blank including both factual and attitudinal questions related to educational and family background. Fourteen multiple choice questions were given unit scoring weights. Some examples of questions are these:

Would you like to go to college if you could afford it? (Check one) — a. Yes; — b. No; — c. Not sure.

Which of the following best describes your High School course of study? (Check one) — a. College preparatory or academic; — b. Commercial and secretarial; — c. General; — d. Other \_\_\_\_\_.

Which of the following occupational groups best describes your father's work during most of his life? (Check one) — a. professional; — b. managerial or executive; — c. own business; — d. clerk in a store; — e. clerk in an office; — f. salesman; — g. skilled trade; — h. farmer or rancher; — i. semi-skilled (factory worker, miner, etc.); — j. Other \_\_\_\_\_; — k. Don't know.

5. The last predictor was a job preference questionnaire which is a modification of the form developed by Jurgensen (3) at the Minneapolis Gas Company. This form requires the respondent to rank 11 factors (Advancement, Benefits, Compensation, Company, Co-workers, Hours, Pay, Security, Supervision, Type of Work, and Working Conditions) in terms of their relative importance to her; and also to rate the importance of having a job which is interesting, important, not strenuous, free from work pressure, uses one's abilities, has much responsibility, and allows freedom for planning one's own work.

### Procedure and Results

This battery was administered to 358 employees in June, 1951. Sixty-five of them left in three months or less and 43 more left from four to twelve months after being hired. Point biserial correlations were computed for each of the five predictor variables for three-month turnover as well as twelve-month turnover. The validity coefficients for three-month and twelve-month turnover are given in Table 1.

Table 1 shows that the General Ability Tests have a validity of — .25 for three-month turnover and — .21 for twelve-month turn-

Table 1

Point Bi-serial Correlations Between Various Predictors and Turnover for Clerical Employees\*

Predictor	3 Month Turnover (Leaving N=65) (Staying N=293)	12 Month Turnover (Leaving N=108) (Staying N=250)
General Ability Tests	-.25	-.21
Clerical Speed Tests	.03	.05
Interest Questionnaire	.19	.19
Biographical Data	.37	.29
Job Preference Blank	.33	.21

\* Negative correlations in this table indicate that those who left scored higher than those who stayed.

over. Negative validity means that girls who left had higher scores than those who stayed. In a previous study of eighteen-month turnover for 1600 girls a validity of  $-.17$  was obtained for these two tests. Clerical speed tests have practically zero validity. In two previous studies these tests had slightly higher validity. The interest turnover key has validity of  $.19$  for both groups. This is a cross-validation result as the key was developed in another study. Biographical Data yields validities of  $.37$  and  $.29$ . Girls who leave, as compared with those who stay, more frequently say they took college preparatory courses, have fathers in professional and managerial jobs, and would like to go to college if they could afford it. Although these identical items have not been used before, similar questions have been used with similar results. The validities of the Job Preference Questionnaire,  $.33$  and  $.21$ , have not been cross-validated. The key for this measure was developed empirically on this sample. Girls who leave, as compared with girls who stay, placed more importance on type of work, pay, and on having a job which used their abilities and gave them freedom to plan their own work. Those who left placed less importance than those who stayed on working for a company they are proud of, on company benefits and on being free from work pressure and strenuous physical requirements. Since these results have not been cross-validated we did not compute intercorrelations between the Job Preference Questionnaire and other predictors and we did not use Job Preference scores in our multiple correlation solution. We will be interested in future cross-validation of the results obtained with this questionnaire.

Intercorrelations among the four variables used in doing multiple correlations are given

in Table 2. In Table 3, you will see that a multiple  $R$  of  $.40$  was obtained for three-month turnover and  $.33$  for twelve-month turnover. In both prediction equations, Biographical Data has much more weight than the other predictors.

Table 2

Intercorrelations of Predictor Variables  
( $N = 358$ )

	Clerical Speed Tests	Interest Question- naire	Bio- graphical Data
General Ability Tests	.20	-.32	-.41
Clerical Speed Tests		-.04	-.10
Interest Questionnaire			.39

In order to determine the practical usefulness of the three-month turnover equation, we examined the data to see what would have happened if, at the time of employment, we had rejected the 35 girls out of the total group of 358 who had the lowest scores on

Table 3

Multiple Correlation Data

Turnover Group	Multiple Point Biserial R	Beta Weights*	
3 Month Turnover	.40	Biographical Data	12
		General Ability Tests	-5
		Clerical Speed Tests	3
		Interest Questionnaire	1
12 Month Turnover	.33	Biographical Data	8
		General Ability Tests	-4
		Clerical Speed Tests	3
		Interest Questionnaire	2

\* High positive score indicates that individual is likely to stay.

Table 4

Effectiveness of Three-Month Turnover Battery:  
Actual Behavior of the 35 Girls with Lowest  
Scores as Compared with Re-  
mainder of Sample

	Left in Less than 3 Months	Stayed 3 Months or More	Total
Accepted	42	281	323
Rejected	23	12	35
Total	65	293	358

the three-month turnover battery. Under present labor market conditions we would not want to reject many more than this. We found that if we had rejected these 35 girls we would have rejected 23 girls who would leave in three months or less and only 12 girls who would stay longer than three months. This means that we would have rejected 36% of the total group who would leave in three months and we would have rejected only 4% of those who would stay longer than three months. Thus it appears that we can use our turnover equations to screen out a substantial percentage of girls who would quit the Company very quickly and would not justify their training expense, and at the same time only lose a small percentage of the girls who become useful long time employees.

#### Summary

As a summary of the findings and implications of this study we would like to make the following points.

1. We can predict quick turnover among newly hired girls for routine clerical jobs moderately well using a combination of Biographical Data, an Interest Questionnaire, General Ability Tests, and Clerical Speed Tests. Biographical Data is the best predictor. The other measures increase only slightly the effectiveness of prediction as estimated by multiple correlation.

2. We can predict turnover for girls who leave in less than three months better than for girls who leave in less than twelve months. As you might infer from this, we cannot predict four to twelve month turnover nearly as well as one to three month turnover. One possible explanation of this is that girls who

leave very quickly are more definitely unsuited for their jobs than those who leave later and therefore their turnover is more predictable. Another possible explanation is that a very high proportion of the three-month turnover group go on to college and this kind of turnover may be especially predictable.

3. The use of General Ability tests with negative weights in selecting girls who will be good turnover risks does not conflict with our aptitude batteries used to predict job performance on beginning assignments, since the valid predictors for most of those jobs are tests of clerical ability rather than the Arithmetic Reasoning and Vocabulary tests.

4. Textbooks in industrial psychology (1, p. 248, 2, pp. 313-314, 4, pp. 89, 97) frequently stress the negative relationship between intelligence and the likelihood of a person staying on a routine clerical job, and recommend the use of upper critical scores on intelligence tests for selecting personnel for such jobs. While we did find the same negative relationship, it is interesting to note that in this study other factors such as family and educational background and interests and aspirations tend to be more important than intelligence.

For our purposes we do not think it necessary or desirable to use an upper critical score on intelligence. General ability scores are related to success on most of our higher level jobs, and in order to have girls with potentiality for advancement it is necessary to hire a number with high general ability. Fortunately our research indicates we can hire girls with such ability who will be fairly good turnover risks as well as good performers on beginning jobs if we screen them carefully on biographical and interest measures, and clerical aptitude tests.

Received November 14, 1952.

#### References

1. Bingham, W. V. *Aptitudes and aptitude testing*. New York: Harper, 1937.
2. Burt, H. E. *Principles of employment psychology*. New York: Harper, 1942.
3. Jurgensen, C. E. Selected factors which influence job preferences. *J. appl. Psychol.*, 1947, 31, 553-564.
4. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, 1947.

## Per Cent Increase in Proficiency Resulting from Use of Selective Devices

Clarence W. Brown and Edwin E. Ghiselli

*University of California, Berkeley*

The common way of presenting evidence concerning the validity of a selective device is by means of the validity coefficient. When this coefficient is high then the device is said to be useful as a means for evaluating candidates for a job, and when it is low the device is said to be ineffectual. Taylor and Russell (2) have shown that this notion is too simple to adequately describe the situation, since gains resulting from the use of a selective device also will be a function of the proportions of persons selected and rejected. When the validity coefficient is low and the ratio of the number of persons selected to the number of applicants is low, the gains may be greater than when the validity is high and the selection ratio also high.

Taylor and Russell's approach has been to evaluate the effectiveness of the results of selection in terms of the proportion of selected persons who turn out to be successful on the job. That is, some cut-off point is set on the criterion and all individuals who meet or exceed this critical point are deemed successful, while those falling below are termed unsuccessful. In many situations this approach is exceedingly useful. Thus, in a training program where a specified proportion of persons are to be passed, knowing the validity of a test and the proportion of persons who will be selected, an estimate can be made of the proportion of those selected who will pass the course. Gains through use of the test can then be expressed in terms of the increase in the proportion of persons passing the training program.

In other situations, however, this is not the information desired. Rather, what is wanted is some estimate of the proficiency of those selected as they are measured by some continuous scale. Thus the question might be asked, if a test of known validity is used and a given proportion of candidates is selected on the basis of their scores, how will the output

of the selected workers compare with that of the unselected workers.<sup>1</sup> If the average production of selected workers is not much greater than that of unselected workers, then the test will not be worthwhile even though it possesses high validity. Furthermore, having an estimate of the potential proficiency of the selected workers will make it possible to improve the planning of production schedules. Suppose, for example, it were desired to place on a particular job persons whose average production is a given amount. Knowing the validity of the test, the proportion to be selected to achieve a certain production schedule could be determined.

Jarrett has recently considered this problem and has developed a formulation which permits the appropriate estimates to be made (1). As with the Taylor-Russell approach normal linear correlations are assumed. The data necessary to estimate gains in proficiency from use of a selective device are the validity coefficient, the proportion of cases to be selected, and if per cent gains are to be estimated, the mean and standard deviation of the criterion scores of unselected cases.

Table 1 is the basic table that has been developed from Jarrett's formulation. This table gives the mean of the standard criterion scores of the selected cases in relation to the validity and the selection ratio. The basic distribution of standard scores is of the unselected cases, and has a mean of zero and a standard deviation of unity. For example, suppose the validity of the selective device is .50 and the 25% highest scoring candidates are selected, then the mean criterion score of

<sup>1</sup> As used in this paper the term "unselected" will have the same meaning as given in Taylor and Russell's (2) and Jarrett's (1) discussions. It will refer to "the members of that population of individuals to who apply for the job in question and who—when individuals are needed for the job in question—would have been put to work without further regard for their qualifications before the testing program was initiated."

Table 1

Mean Standard Criterion Score of Selected Cases in Relation to Validity and the Selection Ratio

		Validity Coefficient																				
		.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
Per Cent Selected on Basis of Test	5%	.00	.10	.21	.31	.42	.52	.62	.73	.83	.94	1.04	1.14	1.25	1.35	1.46	1.56	1.66	1.77	1.87	1.98	2.08
	10	.00	.09	.18	.26	.35	.44	.53	.62	.70	.79	.88	.97	1.05	1.14	1.23	1.32	1.41	1.49	1.58	1.67	1.76
	15	.00	.08	.15	.23	.31	.39	.46	.54	.62	.70	.77	.85	.93	1.01	1.08	1.16	1.24	1.32	1.39	1.47	1.55
	20	.00	.07	.14	.21	.28	.35	.42	.49	.56	.63	.70	.77	.84	.91	.98	1.05	1.12	1.19	1.26	1.33	1.40
	25	.00	.06	.13	.19	.25	.32	.38	.44	.51	.57	.63	.70	.76	.82	.89	.95	1.01	1.08	1.14	1.20	1.27
	30	.00	.06	.12	.17	.23	.29	.35	.40	.46	.52	.58	.64	.69	.75	.81	.87	.92	.98	1.04	1.10	1.16
	35	.00	.05	.11	.16	.21	.26	.32	.37	.42	.48	.53	.58	.63	.69	.74	.79	.84	.90	.95	1.00	1.06
	40	.00	.05	.10	.15	.19	.24	.29	.34	.39	.44	.48	.53	.58	.63	.68	.73	.77	.82	.87	.92	.97
	45	.00	.04	.09	.13	.18	.22	.26	.31	.35	.40	.44	.48	.53	.57	.62	.66	.70	.75	.79	.84	.88
	50	.00	.04	.08	.12	.16	.20	.24	.28	.32	.36	.40	.44	.48	.52	.56	.60	.64	.68	.72	.76	.80
	55	.00	.04	.07	.11	.14	.18	.22	.25	.29	.32	.36	.40	.43	.47	.50	.54	.58	.61	.65	.68	.72
	60	.00	.03	.06	.10	.13	.16	.19	.23	.26	.29	.32	.35	.39	.42	.45	.48	.52	.55	.58	.61	.64
65	.00	.03	.06	.09	.11	.14	.17	.20	.23	.26	.28	.31	.34	.37	.40	.43	.46	.48	.51	.54	.57	
70	.00	.02	.05	.07	.10	.12	.15	.17	.20	.22	.25	.27	.30	.32	.35	.37	.40	.42	.45	.47	.50	
75	.00	.02	.04	.06	.08	.11	.13	.15	.17	.19	.21	.23	.25	.27	.30	.32	.33	.36	.38	.40	.42	
80	.00	.02	.04	.05	.07	.09	.11	.12	.14	.16	.18	.19	.21	.22	.25	.26	.28	.30	.32	.33	.35	
85	.00	.01	.03	.04	.05	.07	.08	.10	.11	.12	.14	.15	.16	.18	.19	.20	.22	.23	.25	.26	.27	
90	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	
95	.00	.01	.01	.02	.02	.03	.03	.04	.04	.05	.05	.06	.07	.07	.08	.08	.09	.09	.10	.10	.11	

the selected cases would be .63 standard deviations above the mean criterion score of the unselected cases. By reversing signs, the mean criterion score of rejected cases can also be estimated. In the case just given, the mean criterion score of the rejected 75% of cases would be .21 standard deviations below the mean of unselected cases.

It is apparent from Table 1 that the smaller the selection ratio is the greater will be the mean criterion performance of the selected cases. Reduction in the selection ratio results in an increase in mean criterion scores, the relationship being positively accelerated, with selection rates smaller than about 20% to 30%. Similarly, as validity increases there is an increase in the mean criterion score of the selected cases. In this instance, however, it will be noted that gains are directly proportional to increase in validity.

In many cases the interest will not be in the standard criterion scores of the selected group but rather in raw criterion scores. Knowing the mean standard score of the selected group, and the mean and standard de-

viation of the unselected group, the desired transformation, of course, can easily be made. Thus in the case already given where the mean standard score of the selected cases was .63, if the mean and standard deviation of the raw criterion scores of the unselected cases were 50 and 10 respectively, the mean raw criterion score of the selected cases would be 56.3. The per cent improvement in proficiency through selection, therefore, would be 12.6.

The appropriate calculations have been performed for various values of the ratio  $\sigma/M$  and are presented graphically in Figure 1. An example will illustrate how this chart is read. Suppose we have a test with a validity of .50 and we are planning to select 20% of persons earning highest scores, the ratio of the standard deviation to the mean of the raw criterion scores of the unselected group being .2. Locating the value of the per cent selected, that is 20%, at the bottom and left of the chart, the line is followed up until it intersects with the curve representing a validity of .50. Now we follow the line across to the right until it intersects with the vertical

line representing the  $\sigma/M$  ratio of .2. Per cent improvement is determined from the placement of this point in the series of curves; in the present case this value would be interpolated as approximately 14%. The mean standard score of the selected group, as also read from the center column of the chart (standard criterion scores of selected cases), is .7.

The chart, of course, can also be read in the reverse direction. Suppose the ratio of the

standard deviation to the mean of the raw criterion scores of the unselected cases is .25 and it is desired to improve criterion performance by 20% through selection of personnel. Locating the point at the intersection of the vertical line for a  $\sigma/M$  of .25 and the curve for 20% improvement, following a line horizontally will give various values of validity and of selection ratio that will produce the desired result. If as many as 50% of applicants are to be selected, then test validity

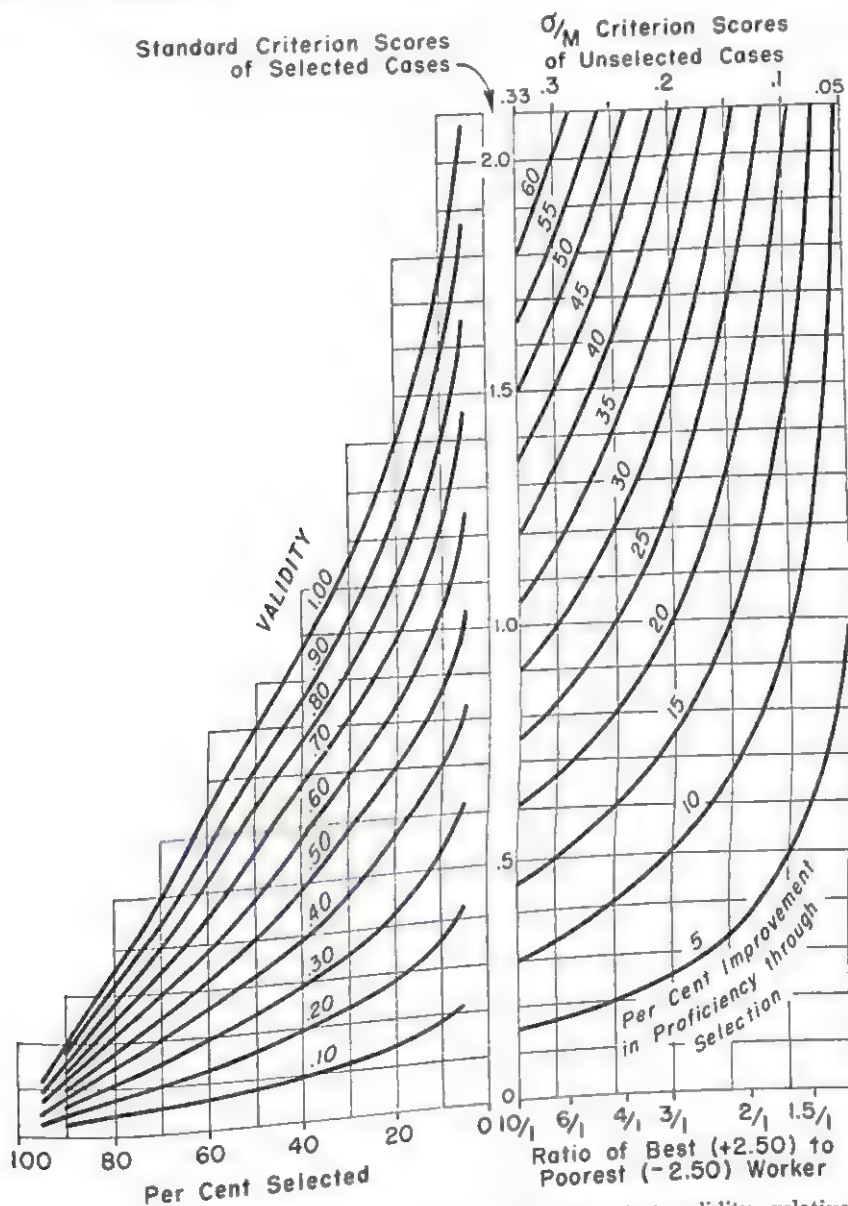


FIG. 1. Interrelationships among the selection ratio, test validity, relative variation in criterion performance, and per cent improvement in proficiency through selection.

would have to be perfect (1.00). With a more reasonable validity of .40, only the best 5% could be selected.

It is apparent from Figure 1 that as the unselected workers become more homogeneous in their criterion performance, that is, as the value of  $\sigma/M$  decreases, the smaller will be the gain from the selection device. For example, with a validity of .50 and a selection ratio of 10%, if  $\sigma/M$  is .3 then improvement will be of the order of 27%. However, if the  $\sigma/M$  is .05 then the per cent improvement will only be about 5%. Probably the limiting case for heterogeneity of criterion performance can be taken as a  $\sigma/M$  of .33, the standard deviation being one third the magnitude of the mean. Since sometimes heterogeneity of criterion performance is expressed in terms of the ratio of the output of the best to that of the poorest worker, an appropriate scale for such values is given on the chart at the foot of the right half of the figure. Since values here must be chosen arbitrarily, performance of the best worker is taken as being + 2.5 standard deviations in the distribution of criterion scores and the poorest as - 2.5 standard deviations.

The picture of the value of selection as given by this approach is by no means too favorable. A validity of .50 is about as high as can be expected in most instances and seldom can a selection ratio be less than 10%. A generous value of  $\sigma/M$  would be .25 (ratio of best to poorest worker being 4 to 1). For these values it will be seen from Figure 1 that the expected improvement in criterion performance is only 23%. In most cases validity will be somewhat lower, the selection ratio higher, and criterion performance more homogeneous. Under optimal conditions, therefore, improvement in productivity as a result of a selection program can be considered to approximate 25%.

*Received November 10, 1952.*

#### References

1. Jarrett, R. F. Per cent increase in output of selected personnel as an index of test efficiency. *J. appl Psychol.*, 1948, 32, 135-145.
2. Taylor, H. C. and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: tables and discussion. *J. appl. Psychol.*, 1939, 23, 565-578.

# Efficiency of Tests When Used to Select the Better of Two Workers<sup>1</sup>

Laurence Kashdan

Civilian Personnel Research Branch, United States Air Force, Washington, D. C.

In most personnel selection situations where tests are used other sources of information about applicants are considered as well in deciding whether to accept or reject people. Most writers of textbooks in employment psychology recommend the use of tests in just this way—as supplements to other valid personal data. Used in this way any test, no matter what its validity, may vary considerably as to the role it will play in a given company's program, or among different companies.

How much weight should a personnel officer place upon the test scores of two people when other data about them are also available? Even low validity coefficients become actuarially significant in the course of many decisions based upon test scores.<sup>2</sup> The Taylor-

Russell tables<sup>3</sup> specify the efficiency of selection using tests whose validity coefficients may vary between 0 and 1, depending upon various existing employment conditions. However, these tables assume that all decisions will be made over the long term on the basis of test scores alone.

What is needed is a guide which will help a personnel officer decide about the risk he may be assuming in relying entirely upon the achievement in a test by two or more applicants for a position—or in choosing to ignore relative scores in favor of non-test considerations. It is possible to specify the probability that a test has correctly ranked two people in terms of a criterion of job performance when the difference in their standard scores on the test is known. If the assumptions for computing the product-moment coefficient of correlation had been properly met in computing the validity coefficient, and all scores are now

<sup>1</sup> The author is very grateful to Professor James N. Mosel of George Washington University, discussions with whom suggested the main concepts of this paper. It is also a pleasure to thank Dr. John R. Boulger of this office for his helpful review of the manuscript.

<sup>2</sup> Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, 1952.

<sup>3</sup> Taylor, H. C. and Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *J. appl. Psychol.*, 1939, 23, 565-578.

Table 1  
The Probability of Selecting the Better of Two Workers on the Basis of the Difference in Their Test Scores

Difference Between Two Test Scores in Standard Score Units	Validity Coefficient of the Test ( <i>r</i> )								
	.1	.2	.3	.4	.5	.6	.7	.8	.9
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50
.25	.50	.52	.52	.53	.53	.55	.57	.59	.64
.50	.51	.53	.54	.56	.58	.61	.64	.68	.77
.75	.52	.54	.57	.59	.62	.66	.70	.76	.86
1.00	.53	.56	.59	.62	.66	.70	.75	.83	.93
1.25	.54	.57	.61	.65	.70	.75	.81	.88	.97
1.50	.54	.59	.63	.68	.73	.79	.85	.92	.99
1.75	.54	.59	.63	.68	.73	.79	.85	.92	.99
2.00	.55	.60	.65	.70	.76	.82	.89	.95	1.00
2.25	.56	.61	.67	.73	.79	.84	.92	.97	1.00
2.50	.56	.63	.69	.75	.82	.88	.94	.98	1.00
2.75	.57	.64	.71	.78	.85	.91	.96	.99	1.00
3.00	.58	.66	.73	.80	.87	.93	.97	.99	1.00
	.58	.67	.75	.82	.89	.94	.98	1.00	1.00

expressed as standard scores, then we can derive Table 1.<sup>4</sup> This gives the probability that for any two subjects selected at random from among those taking the test, the one of them who has earned the higher score in the test will be the better worker—in terms of the criterion of validity for that test.

As an example of how Table 1 would be applied in a practical situation consider the following data:

A achieves a standard score of .95 in a test; B earns a score of .20; the validity coefficient of the test is .7; what is the probability that A will prove to be a better worker than B?

<sup>4</sup> See: Jenkins, W. L. An index of selective efficiency (S) for evaluating a selection plan. *J. appl. Psychol.*, 1953, 37, 78, for a comparable treatment which disregards test score difference.

Since the difference between the scores of the two men is .75 standard score units, the first column of the table is entered at .75, and moving over to the column for  $r = .7$  the tabled probability is given as .70. This means that on the basis of test score alone there are about 7 chances in 10 that A will turn out to be the better worker. Or, viewed conversely, if the personnel officer should decide to disregard their relative achievement on the test and select B over A for the job, there would be only about 3 chances in 10 that his decision will prove to be correct.

Table 1 is an effective way to illustrate the meaning of a validity coefficient to personnel people in terms of their own operations.

*Received July 6, 1953.*

*Early publication.*

## Ratings of Candidates for Promotion by Co-workers and Supervisors

Doris Springer

*Supervisory Selection Board, North American Aviation, Inc., Los Angeles, California*

The primary purpose of this study was to compare ratings made by supervisory personnel and by co-workers on candidates for promotion to leadman jobs. Specifically, answers to the following questions were sought:

- (1) To what extent do supervisory personnel and co-workers agree in their ratings of workers?
- (2) How does the extent of this agreement compare with (a) the extent to which members of supervision agree with each other, and (b) the extent to which co-workers agree with each other in the ratings given workers?

In analyzing the data to answer these questions, answers were suggested for other questions, such as:

- (3) How do judgments on different items in the rating form compare with each other?
- (4) Is there any evidence that supervisors tend to rate candidates lower or higher than do their co-workers?
- (5) How do the totals of the ratings on the individual characteristics compare with the ratings on the suitability of the candidate for promotion?

The problem is of practical importance in determining the reliability of ratings by supervisors and co-workers and in arriving at the appropriate weights to be given ratings made by them in an over-all evaluation of candidates for promotion. The provision in many union contracts which states that promotions to jobs covered by the contract are to be governed by seniority only when ability, skill and job performance are equal draws attention to the need for devising techniques for determining workers' suitability for promotion. These techniques must be acceptable to the union and management and, at the same time, be statistically sound. It then becomes important to analyze the results of these techniques in their actual application. The pres-

ent study provides data on two of these techniques, namely, co-worker ratings and supervisory ratings.

From a theoretical standpoint, the study contributes some data on the attitudes of two distinct groups in the economic structure and on the relative homogeneity of thought of these two groups with respect to one aspect of their work environment. An accumulation of such data will enable us at some future time to arrive at a psychological and sociological understanding of the two groups which will be invaluable to the industrial psychologist.

### Ratings Studied

This study is based on the ratings made on 100 men who were candidates for leadman<sup>1</sup> jobs in 14 different departments of the manufacturing division of a major aircraft company. The ratings were made as a regular phase of the company's supervisory selection program in which each candidate is evaluated on the basis of his work experience, education, work record, and scores on mental ability, shop math, and job knowledge tests, in addition to the ratings. Ratings are made by two supervisors, representing two levels of supervision over the candidate, and by three co-workers who work closely with the candidate but who are not eligible to be candidates for the leadman job. The ratings analyzed here are the ratings made by two members of supervision and two of three co-workers (selected at random) for each of the 100 candidates. A total of 68 different assistant foremen and foremen made the supervisory ratings. The exact number of co-workers participating cannot be reported since these rating forms were not signed, but the number was probably between 150 and 175.

<sup>1</sup> At North American Aviation, Inc., a leadman directs a group of five to ten men. The job is covered by union contract.

A worker ordinarily rated only one worker for any one job opening and rarely did a job opening occur in the same group during the period studied.

The two rating forms used were the "behavior sample" type in which five gradations from very poor to outstanding were described for each characteristic. The form used by the co-workers consisted of five factors; namely, job knowledge, job performance, co-operation, ability to train others, and suitability for promotion to leadman. The form used by supervisory personnel consisted of eight factors; namely, job knowledge, quality of work done, quantity of work done, co-operation, drive, observing rules, personal appearance and manner, and suitability for promotion to leadman. The raters were instructed to check the one statement for each factor which best described the candidate. The ratings were made independently. For purposes of this report, the five intervals have been assigned values of 1 through 5, from lowest to highest.

#### Statistical Method

The degree of relationship between the variables studied has been measured by the product moment coefficient of correlation. It was believed that the nature of the data justified the use of this technique because the series were more nearly continuous than discrete and more nearly quantitative than qualitative.

When a difference is described in the report as significant, that difference is so large that it could be expected by chance not more than once in 100 times ( $P \leq 0.01$ ).

#### Results

*Relationship between ratings made by members of supervision and co-workers.* The coefficients of correlation between the ratings made by one member of supervision and one co-worker for each candidate on items common to both rating scales are shown in Table 1. The single supervisory rating was chosen at random from the two ratings made and the single co-worker rating was chosen at random from the three ratings made. Supervisory ratings on quality of job performance and quan-

tity of work done have been compared with the single rating on job performance given by co-workers.

All of the correlations are rather low, ranging from .15 to .39; however, only the lowest coefficient is not significantly greater than zero. There is greatest agreement on the over-all rating of general fitness for promotion. The data in Table 1 suggest that co-

Table 1  
Relationship Between Ratings Made by Members  
of Supervision and Co-workers

Item Rated	Coefficient of Correlation
Job knowledge	.15
Job performance—Quality	.25
Cooperation	.29
Job performance—Quantity	.33
General fitness for promotion	.39

worker and supervisory ratings do not duplicate each other unnecessarily; and, *at least in this respect*, the consideration of both types of ratings in evaluating candidates for promotion seems justified.

The low degree of agreement between the ratings of supervisory personnel and co-workers indicates that many factors determining the ratings of the two groups are either not similar, or are not receiving the same relative emphasis. Perhaps their standards of judgment, based on differences in scope and type of experience and present job status, account for the lack of agreement. Their ratings may be determined by observations of different samples of behavior of the men being rated. On the other hand, the discrepancies in the ratings found here may be accounted for, in part, by differences of opinion on what characteristics are desired in a leader of the work group. Research on worker and supervisor attitudes with regard to how work groups should be led has suggested that such differences exist (1).

The data reported here merely show that differences between the opinions of co-workers and members of supervision do exist; further research is necessary to identify the sources of these differences.

*Relationship between ratings made by pairs of co-workers.* The coefficients of correlation between ratings made by pairs of co-workers on the candidates are shown in Table 2. The coefficients indicate an agreement between pairs of co-workers which, although greater than zero, is moderately low to moderate.

Table 2

Relationship Between Ratings Made by  
Pairs of Co-workers

Item Rated	Coefficient of Correlation
Cooperation	.34
General fitness for promotion	.34
Instruction ability	.41
Job performance	.41
Job knowledge	.43
Total of all items	.48

With one exception, the correlations between ratings made by pairs of co-workers are higher than the correlations between co-workers and supervisory personnel. There is slightly less agreement among co-workers than between co-workers and supervisors on general fitness for promotion; however, the difference is not significant.

When ratings given for all items on the rating form are combined, the coefficient obtained is slightly higher (though not significantly so) than for any individual item. The coefficients in Table 2 are in line with those reported in most studies of supervisory merit ratings (2, 3, 4).

The greater agreement among co-workers than between co-workers and supervisors may reflect more similarity among the former than between the latter with respect to standards of judgment, behavior actually observed, and/or opinions on what characteristics are desired in a leader of the work group.

The fact that only moderate agreement is found indicates that the co-workers are far from being a homogeneous group with respect to attitudes toward their co-workers.

The comparison here may be interpreted as a measure of the reliability of the co-worker ratings. The moderately low to moderate reliability of the ratings indicates

that such ratings should not be used as the sole basis for selection and that care must be taken in their interpretation. The relatively low reliability of co-worker ratings, as compared with reliability coefficients of other types of measures, should be considered in deciding on the weight of these ratings in the battery of measurements to be used in evaluating the candidates.

*Relationship between ratings made by pairs of supervisory personnel.* The coefficients of correlation between the ratings made on each candidate by two members of supervision are shown in Table 3.

The coefficients, ranging from .56 to .71, indicate a fairly high degree of agreement between the members of supervision in rating workers on all items included in the rating scale. The over-all rating, general fitness for promotion, showed the highest degree of agreement although none of the differences between the items are clearly significant. The fairly high correlations indicate that members of supervision tend to base their ratings on similar observations of the workers' performance and to judge the various characteristics according to similar standards.

All of the coefficients reported in Table 3 exceed those reported in the previous comparisons and suggest a greater degree of agreement among members of supervision than among co-workers and between co-workers and members of supervision.

If the relationship is interpreted as a measure of reliability, then the supervisory ratings

Table 3

Relationship Between Ratings Made by  
Pairs of Supervisors

Item Rated	Coefficient of Correlation
Observing rules	.56
Personal appearance	.61
Quality of work	.61
Job knowledge	.63
Drive	.65
Quantity of work	.66
Cooperation	.67
General fitness for promotion	.71
Total of all items	.66

Table 4  
Distributions of Ratings by Supervisors and Co-workers

Item Rated and Rater	Number of Ratings in Interval					Mean of Ratings	Standard Deviation
	1	2	3	4	5		
Job knowledge							
Supervisors	1	8	73	80	38	3.73	.83
Co-workers	1	8	55	63	73	4.00	.92
Job performance—Quantity*							
Supervisors	0	5	84	68	43	3.74	.82
Co-workers	1	6	44	76	73	4.07	.86
Job performance—Quality							
Supervisors	0	1	56	86	57	4.00	.76
Co-workers	1	6	44	76	73	4.07	.86
Cooperation							
Supervisors	0	8	90	44	58	3.76	.92
Co-workers	1	10	41	67	81	4.08	.94
General fitness for promotion							
Supervisors	3	35	61	60	41	3.50	1.05
Co-workers	3	16	50	65	66	3.87	1.01

\* Supervisory ratings on quality of work done and quantity of work done are compared with co-worker ratings on job performance which included both quality and quantity.

have a fairly high degree of reliability. The greater consistency of the supervisory ratings as compared with co-worker ratings suggests that the former are more dependable.

*Comparison of the distributions of ratings by members of supervision and co-workers.* The distributions of the ratings by the 200 members of supervision and the 200 co-workers on the items common to both rating forms are shown in Table 4.

The ratings of supervisors tend to be more conservative than those of the co-workers. This is evident in a comparison of the proportions of ratings of the two groups which are in the highest interval in the rating scale (step 5). For every characteristic rated a smaller proportion of the supervisory ratings is in the top interval than is true of co-worker ratings. In only one instance is the difference small enough to be attributed to chance (for job performance-quality,  $P = .10$ ).

The tendency of supervisors to give lower ratings than co-workers is shown also in a comparison of the means of the various items rated. In every instance the mean of the supervisory ratings is lower than the mean of the co-worker ratings. The differences are significant at the 1% level, or better, with the exception of job performance-quality, where  $P = .19$ .

Very few of the workers were rated in the lowest category by either supervisors or co-workers. Since there had been some prior selection of the men (they had been proposed for consideration by either members of supervision or of Industrial Relations), it was expected that seldom would a candidate be rated as very unsatisfactory in any factor. Although the frequencies in the second interval are higher than in the lowest interval, the second interval is used in fewer than 5% of the ratings except for the over-all rating. For the item, general fitness for promotion, approximately 18% of the ratings of supervisory personnel and about 8% of the ratings of co-workers are in the next to the lowest interval.

The interval with the highest total frequency is the third, or average, interval for members of supervision and the fifth, or top, interval for co-workers. In the case of supervisors, for three of the five factors shown, the modal interval is step three, and for the other two factors is the fourth interval. For co-workers, step five is the modal interval for three of the four different factors shown. Thus, another method of analyzing the data shows that members of supervision tend to give lower ratings than do co-workers.

Several explanations might be suggested

for the relatively low ratings given by members of supervision as compared with co-workers. Perhaps the status of supervisory personnel results in more realistic, less personal ratings. Also, members of supervision have had more training in the use of the rating form since many of them attend meetings of the Supervisory Selection Board. Some of them had reviewed the rating forms when the forms were being constructed.

*Comparison of the individual items on the rating forms.* In a comparison of ratings assigned to the various items shown in Table 4, it appears that the distribution for the final over-all ratings on suitability for promotion differs from the distributions on the other factors of ratings by both supervisors and co-workers. For example, the mean of the ratings for this factor is significantly lower than the mean of the ratings assigned any of the other items. A greater proportion of the ratings on this factor are in the two lowest intervals (below average) than is true of any other factor; however, only in the case of the supervisory ratings are the differences clearly significant.

The differences between the standard deviations for the ratings given by the two groups of raters are not statistically significant. The greatest variation in ratings of both groups is found in the ratings on general fitness for promotion.

When the final item, suitability for promotion to leadman, is compared with the total of the ratings on all other items in the rating forms, the correlations obtained are .85 for supervisors and .85 for co-workers. The coefficients approximate the ones reported in previous studies in which the same type of comparison was made (3, 4).

### Summary and Conclusions

A group of 100 men who were candidates for promotion to leadman jobs in the manufacturing division of an aircraft company were rated by members of supervision and by co-workers. Comparisons were made between ratings given each candidate by: (1) a member of supervision and a co-worker; (2) two members of supervision; and (3) two co-workers. The following conclusions are based on the results of these comparisons:

1. There is a low, positive degree of relationship between the ratings given by supervisory personnel and co-workers.

2. There is a slightly higher degree of agreement between the ratings of pairs of co-workers than between the ratings of members of supervision and co-workers. The correlations obtained indicate a moderately low to moderate statistical reliability for the co-worker ratings.

3. There is a much higher degree of agreement among the ratings given by members of supervision than among ratings given by co-workers. The correlations obtained indicate a fairly high statistical reliability for the supervisory ratings.

4. Supervisory personnel tend to rate the men lower than do co-workers on all items common to the two rating forms as shown by consistently lower mean ratings, by lower modal intervals, and by a larger proportion of candidates considered below average on general fitness for promotion.

5. Both members of supervision and co-workers tend to be somewhat more conservative when rating the candidates on the over-all item, general fitness for promotion to leadman, than when rating individual characteristics.

6. There is a very high degree of relationship between the total of ratings on all separate characteristics and the ratings given on the single item, general fitness for promotion.

*Received July 20, 1953.*

*Early publication.*

### References

1. Fleishman, E. A. The measurement of leadership attitudes in industry. *J. appl. Psychol.*, 1953, 37, 153-158.
2. Ghiselli, E. E. The use of the Strong Vocational Interest Blank and the Pressey Senior Classification Test in the selection of casualty insurance agents. *J. appl. Psychol.*, 1942, 26, 793-799.
3. Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book, 1940, pp. 49-72.
4. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, Inc., 1952, pp. 345-346.
5. Williams, S. B. and Leavitt, H. J. Group opinion as a prediction of military leadership. *J. consult. Psychol.*, 1947, 11, 283-291.

## Turnover Factors as Assessed by the Exit Interview

Frank J. Smith and Willard A. Kerr

*Illinois Institute of Technology*

Many employees in the process of quitting their jobs are in a mood to express feeling and speak frankly. If the enterprise maintains a formal exit interview in which the employee is assured that nothing he says will be used "against him" in any way, the tendency toward frankness and even catharsis is strengthened. The *exit interviewer* thus is in a uniquely advantageous position to observe the dynamics of the turnover process. Different approaches (1, 5, 8, 10, 11) to study of turnover are desirable; undoubtedly avoidable turnover differs from one enterprise to another in qualitative ways because of differing organizational climates and the patterns or syndromes of reasons for quitting should in part be products of these climates.

### Experimental Design

On the assailable but necessary assumption that exit interviewers are adequate media for assessing the patterns of turnover, the following research was executed. A brief content analysis report<sup>1</sup> was constructed, requesting the exit interviewer to estimate how often in five typical interviews each of sixteen topics was "mentioned as a reason for leaving." This report form with a cover letter was sent to the exit interviewer in each of 200 different, nationally representative companies (selected randomly from *Poor's*). Of these, nineteen replied that they did no exit interviewing and two were returned unclaimed. Another two were returned with verbal explanations but without usable quantitative data. Forty-eight properly completed analyses were returned and utilized in this research. The 48 companies are geographically representative

and report an annual exit interview case load of 5075.

*Instrument Reliability.* A split-half reliability coefficient for the content analysis report on the 48 returns was .81 which became .90 when corrected by the Spearman-Brown formula.

### Results

*Exit Interview Content Profile.* According to reports from these 48 companies, each topic is mentioned as a reason for leaving as follows (per five representative interviews): pay, 1.89; transportation, 0.81; promotion, 0.73; working conditions, 0.69; poor health, 0.64; job security, 0.54; friction with co-workers, 0.52; poor housing or excessive rents, 0.50; personal happiness as affected by job experience, 0.33; ability of supervisor, 0.33; broken promises by supervisor, 0.25; confidence in management, 0.19; company interest in employee welfare, 0.15; freedom of communication with higher levels, 0.12; recreation, 0.04; method of wage payment, 0.02; other problems, 1.15.

*Comparison of Content Profile with Job Satisfaction Data.* Most of the above topics are included in a widely used job satisfaction survey form (7). When some typical survey results (9) were compared with these exit interview content data, it was found that pay was the foremost grievance in both. Working conditions also was a major grievance in both. Among other topics, however, the agreement was moderate or low.

The results just quoted which emphasize employee concern about pay may seem to contradict some previous research. It is true that many researchers (2, 3, 4, 6, 12, and others) have found that when employees are asked what they consider "most important" in their jobs, they do not put pay as foremost in importance. Actually, such results are not in contradiction to the job satisfaction survey and exit interview results, because the "importance ranking" studies represent research

<sup>1</sup> The questionnaire, copies of cover letters, a copy of the exit content profile, the correlation matrix, and the job satisfaction data used in comparison have been deposited with the American Documentation Institute. Order document No. 4054 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting \$1.25 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.25 for photoprint readable without optical aid.

on an entirely different variable. "Factor importance in a job" is not the same thing as what the employee is happy or unhappy about in a job. The factor importance ranking studies referred to above are cast in an abstract, theoretical frame of reference for the employee respondent. They get at his set of philosophical values. But the job satisfaction survey and exit interview get at something different: *not the importance but the satisfactory or unsatisfactory condition of each factor in current work experience*. Generalizing from all these related researches, we might suggest that employees in general concede that pay is not the most *important* factor in a job, but

they nevertheless feel that it represents a foremost *grievance* factor. This generalization is bolstered by a non-attitudinal turnover study, revealing pay as a foremost objective correlate of turnover (8).

*Comparison of Exit Content Profile with Routine Personnel Counseling Profile.* A content analysis report form substantially identical to the one used for exit data except that it was focused upon routine personnel counseling was constructed and sent to 39 companies which at some time had had counseling programs. The eight firms which finally co-operated returned reports from 22 personnel counselors who reported serving an annual

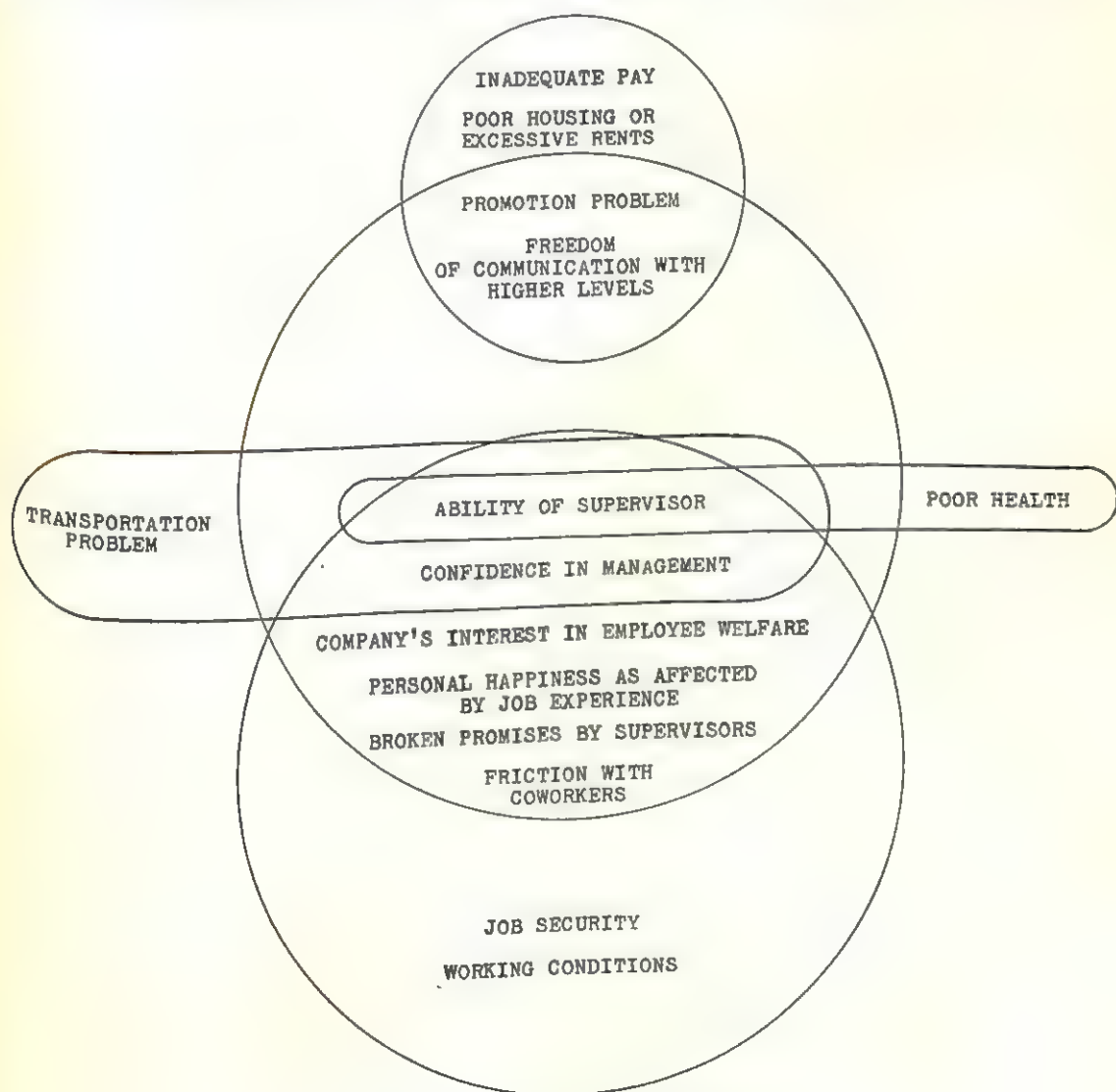


FIG. 1. The exit interview content patterns.

total load of approximately 30,000 cases. A sixteen topic profile was computed on these 22 reports. The split-half reliability was .85 which corrected to .92. When these profile values were correlated with the similarly-derived exit interview profile values  $\rho$  was found to be .74. Apparently the content of routine personnel counseling interviews is very similar to the content of the exit interview.

*Pattern Structure of Exit Interview Content.* Intercorrelations were computed among the sixteen exit interview content topic frequencies for the 48 companies. A simple linkage-type cluster analysis of the resulting matrix was then performed in an effort to isolate the most characteristic exit patterns or climates. Results of this analysis are shown schematically in Figure 1.

Perhaps the most conspicuous single outcome of the analysis is the presence in four of the five clusters of "ability of the supervisor." Even "poor health," which correlates with nothing else as a reason for quitting, correlates .75 with (lack of) "ability of supervisor." The question of whether the supervisor is a convenient scapegoat for the employee in poor health or whether he bears a causal psychosomatic relationship to employee poor health is not answered in these data. The triad in which the supervisor also figures along with "transportation" grievances and lack of "confidence in management" suggests an interesting pattern in some companies. It appears probable that employees living far away from the plant have more transportation difficulties and therefore are tardy or absent more frequently than other personnel. The supervisor (of this pattern) categorizes mentally and orients to these employees as being of the "less dependable, tardy, absentee type." Gradually the distant-living employee perceives this apparent untrusting attitude in the supervisor, and he develops a reciprocal lack of confidence in the management. Eventually, according to this plausible interpretation, he ends up in the exit interview complaining about transportation, the ability of the supervisor, and lack of confidence in management.

Three other factors, climates, or syndromes appear. A general *human relations pattern*

emphasizes concern with broken promises by supervision, friction with co-workers, company interest in employee welfare, freedom of communication, promotion, ability of supervisor, confidence in management, and job effect on personal happiness. A *security pattern* emphasizes complaints about job security, working conditions, confidence in management, ability of supervisor, interest in employee welfare, broken promises by supervisor, job effect on personal happiness, and friction with co-workers.

An *upgrade pattern* is evident in a tendency toward simultaneous complaint about promotion, pay, and freedom of communication. Poor housing complaint is also included, correlating negatively with pay complaint but positively with grievances about promotion and freedom of communication. The "rush to get ahead" syndrome is apparent here. Promotion and access to "higher ups" are considered important—along with pay or satisfying family housing. They do not feel that the present employment permits them to meet their pay and prestige goals fast enough.

In interpretation it should be noted that these patterns may, in fact, represent the turnover-inducing climates operating in parts of the 48 enterprises reported on. Unfortunately, to an unknown degree, it is possible that the patterns may be influenced by frames of perceptual reference of the interviewers themselves. Even allowing for some common sets among interviewers, it still seems probable that their reports must also have been influenced by what they have seen, heard, and sensed in their daily work of exit interviewing.

### Summary

Forty-eight exit interviewers in as many companies supplied topical analyses of exit interview content. These data, ostensibly products of differing turnover climates, were summarized by topic, intercorrelated, and analyzed to suggest the following conclusions.

1. Pay grievances were mentioned twice as frequently as any other single topic of complaint. Next in order of complaint were transportation, promotion, working conditions, poor health, job security, co-workers, housing, the job, supervisor, confidence in manage-

ment, interest in employee welfare, freedom of communication with higher levels, recreation, and method of wage payment.

2. The relatively heavy emphasis upon pay and working conditions agrees with the heavy emphasis assigned by regular employees themselves in job satisfaction surveys, and with turnover correlates, but disagrees with "factor importance ranking" studies. Otherwise, exit interview topic emphasis agrees only moderately with "per cent dissatisfied" on job satisfaction surveys of non-quitting personnel.

3. When 22 regular (not exit) personnel counselors submitted reports of content of their routine counseling, their mean profile of topic frequencies was found to correlate ( $\rho$ ) .74 with the mean profile obtained on the 48 exit interviewers. Apparently there is much in common among the frustrations expressed by employees who are quitting and by employees still on the job.

4. A cluster analysis of exit topic frequency intercorrelations was performed with the following climatic patterns resulting: a human relations syndrome; a security syndrome; an upgrade syndrome; a transportation-confidence triad; and an unnamed duad.

Received November 24, 1952.

## References

1. Baruch, Dorothy. Why they terminate. *J. consult. Psychol.*, 1944, 8, 35-46.
2. Blum, M. L. and Russ, J. J. A study of employee attitude toward various incentives. *Personnel*, 1942, 19, 438-444.
3. Chant, S. M. Measuring the factors that make a job interesting. *Personnel J.*, 1932, 11, 1-4.
4. Hersey, R. B. Psychology of workers. *Personnel J.*, 1936, 14, 291-296.
5. Ho, C. J. Health and labor turnover in a department store. *Personnel J.*, 1930, 9, 216-221.
6. Jurgensen, C. E. Selected factors which influence job preferences. *J. appl. Psychol.*, 1947, 31, 553-564.
7. Kerr, W. A. *The tear ballot for industry*. Chicago, 90: Psychometric Affiliates, 1944.
8. Kerr, W. A. Labor turnover and its correlates. *J. appl. Psychol.*, 1947, 31, 366-371.
9. Kerr, W. A. and Cramer, R. J. Age group and attrition morale phenomena in industrially employed males. Paper presented to Midwestern Psychological Association, Detroit, Mich., May 5, 1950.
10. Miller, L. R. Why employees leave. *Personnel J.*, 1944, 23, 111-119.
11. Tiffin, J., Parker, B. T., and Habersat, R. W. The analyses of personnel data in relation to turnover on a factory job. *J. appl. Psychol.*, 1947, 31, 615-616.
12. Wyatt, S., Langdon, J. N., and Stock, F. G. Fatigue and boredom in repetitive work. *British Industrial Health Research Board Report No. 77*, 1937, 43-46.

# The Quartile Difference Method of Item Selection<sup>1</sup>

Norman Friedman<sup>2</sup>

*Occupational Research Center, Purdue University*

It is customary in cases where personal data are used as predictors of various criteria of job performance to use as predictors all of the items which show a significant relationship with the criterion. The drawback to this mode of operation lies in the possibility of some of the included items contributing more error in prediction than they do to actual validity.

This phenomenon can best be explained in terms of item and criterion variance. A hypothetical illustration will be given in terms of two items. Item 1 shares 20 per cent of its variance in common with the criterion, includes 50 per cent specific variance and 30 per cent error variance. The second item shares 15 per cent of its variance with the criterion, but 12 of these 15 per cent are in common with the 20 per cent shared by item 1. Further, 50 per cent of the variance for item 2 is specific, and 35 per cent is error variance. By adding item 2 to item 1, three per cent more of the criterion variance is accounted for; but at the same time 35 per cent additional error variance is introduced. Thus, adding item 2 would result in shrinking the validity of item 1.

In a specific situation, then, the problem is one of selecting a number of items from a pool of items so that the selected items give a maximum relationship with the criterion. The Wherry-Doolittle Technique<sup>3</sup> achieves this goal for data where item validities and

inter-relationships can be expressed in terms of coefficients of correlation. For categorical data, however, where item versus criterion relationships are expressed in  $2 \times 2$ ,  $2 \times 3$ ,  $2 \times k$  contingency tables, neither item validities nor item inter-relationships can be expressed in correlational terms (except for those recorded in  $2 \times 2$  tables). Regardless of the data format, the problem of shrinkage remains the same. The quartile difference method proposed here does essentially with categorical data what the Wherry-Doolittle Technique achieves with correlational data.

The mechanics of this method will be illustrated with four application blank items that were found to be related to the tenure of telephone operators at the 10 per cent significance level or better on the basis of analysis with a primary group of 171 operators. Table 1 lists the four items that were considered to be significantly related to tenure, the per cent of high and low criterion cases in each category, and the scoring weights for the various categories. In addition, the value of chi square for each item along with its degrees of freedom and probability level are listed.

## Method

The quartile difference method involves the following steps:

1. Divide the total sample of cases into a primary group and a holdout group. Working with the primary group, compute chi square for each item. Then, compute scoring weights for the various response categories of the items that are significantly related to the criterion. For the illustrative case, these results are presented in Table 1.

2. List the responses of subjects in the holdout group to the items which were considered to be significantly related to the criterion and assign the scoring weights as determined in Step 1 to these responses.<sup>4</sup> For example, if a subject in

<sup>1</sup> This article is based on part of a Ph.D. dissertation done under the direction of Professor E. J. McCormick. The dissertation is on file in the Purdue University library under the title "Personal Data as Predictors of the Job Behavior of Telephone Operators."

<sup>2</sup> The author wishes to express his gratitude to the General Telephone Company of Michigan, Mr. F. E. Norris, President, whose cooperation made this study possible. In this regard, a special word of thanks is due Dr. Melvin Tieszen, formerly Personnel Director of the Company and now affiliated with Booz, Allen and Hamilton, New York.

<sup>3</sup> Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Company, 1940, pp. 253-255.

<sup>4</sup> The step outlined here follows a cross validation procedure with a holdout group of 176 cases. While the item selection technique may be carried out with a single group of employees, it is strongly recommended, if it is at all feasible, that the cross validation procedure be used.

Table 1

The Four Items Related to Tenure, Per Cent of High and Low Tenure Cases in Each Response Category, Scoring Weights, Item Chi Square with its Degrees of Freedom and Probability Level

Item	Categories	% High Tenure	% Low Tenure	Scoring Weights*	Chi Square	D.F.	P.
1. Height-weight ratio					4.931	2	.09
	2.00-2.04	14	8	28			
	1.70-1.99	47	65	4			
	1.45-1.69	39	27	34			
		100	100				
2. Marital status					3.841	1	.05
	single	79	65	36			
	married	21	35	8			
		100	100				
3. When consult physician					4.822	2	.09
	no mention	31	28	28			
	0-9 months	41	57	6			
	9 months +	28	15	35			
		100	100				
4. Education					7.786	2	.02
	below high	21	12	31			
	high grad.	69	63	28			
	above high grad.	10	25	7			
		100	100				

\* The scoring weights were arrived at by subtracting the per cent of low tenure cases from the per cent of high tenure cases for each category and adding a constant, +22, to eliminate negative weights.

the holdout group fell in the 1.70-1.99 height-weight ratio category, was single, had consulted a physician more than nine months prior to the time of application and had a high school education, the scoring weights (taken from Table 1) for this subject would be listed as follows: 4, 36, 35, 28.

3. Select as the first item to be included in the battery that item which demonstrated the highest relationship with the criterion as determined in Step 1. In this case the item selected was number 4 (Education) with a probability level of .02.

4. List the scoring weights for the first selected item in order of magnitude (from high to low) and tally the frequencies of high criterion and low criterion cases in the holdout group at each scoring weight. Split the total distribution of cases at the various scoring weights into an upper quarter, middle half and lower quarter.<sup>5</sup> Then,

<sup>5</sup> If the first selected item is dichotomous, it would be, of course, impossible to split the total distribution of holdout cases at the various scoring weights into high quarter, middle half and low quarter since the cases in the holdout group are tallied at only two scoring weights. In situations such as this, the scoring weights for the first selected item should be

compute the per cent of high criterion cases in the upper quarter- $Q_1$ , middle half- $Q_m$  and lower quarter- $Q_l$ . The difference in per cent of high criterion cases between the upper and lower quartiles,  $Q_1 - Q_l$ , serves as a measure of item, or item combination, discrimination. These computations for item 4 are presented as the zero order of analysis in Table 2.

5. Plot the  $Q_1$ ,  $Q_m$  and  $Q_l$  values obtained from the zero order analysis on a shrinkage chart, Figure 1. The horizontal line at the 60 per cent point represents the per cent of high criterion cases in the holdout group.

6. Combine the scoring weights for the first selected item with the scoring weights for each of the remaining items for every subject in the holdout group. This procedure will result in as many new distributions of scoring weights as there are items to pair with the first selected item. List the combined scoring weights for each pair of

immediately combined with the scoring weights for the remaining items as is indicated in Step 6 below. In other words, the zero order of analysis is bypassed and the researcher goes immediately to the first order of analysis.

items in order of magnitude (from high to low) and tally the frequencies of high criterion and low criterion cases in the holdout group at each scoring weight for each distribution of scoring weights. Once again split the total distribution of cases at the various scoring weights for each item combination into upper quarter, middle half and lower quarter and compute the per cent of high criterion cases for each of these categories. For this first order analysis the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values as computed for each pair of items are entered in the recording sheet, Table 2. The second item to be selected for the battery is that item which, when combined with the first selected item, yields the highest  $Q_1 - Q_4$  value. In this case, the second item to be selected was item 2 (Marital Status) which, with item 4, yielded the highest  $Q_1 - Q_4$  value, namely 52.

7. Plot the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values for the best two items as selected by the first order analysis on the shrinkage chart, Figure 1. In this case, the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values plotted for the first order analysis were the values obtained for items 4 and 2 from Table 2. For the illustrative case examination of Figure 1, the shrinkage chart, at this point reveals that the addition of item 2 increases the validity of the composite as compared with the validity of item 4 alone (the distance between the  $Q_1$  and  $Q_4$  values continues to spread, indicating increased efficiency in prediction or item combination validity). Consequently, the item selection procedure is continued.

8. Combine the composite scoring weights for the first two selected items with the scoring weights for each remaining item. Once again follow the computational procedure outlined in steps 4 and 6 above. For this second order analysis, the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values are computed for each item triad and recorded in Table 2. The third item to be selected for the battery is that item which, when combined with the first two selected

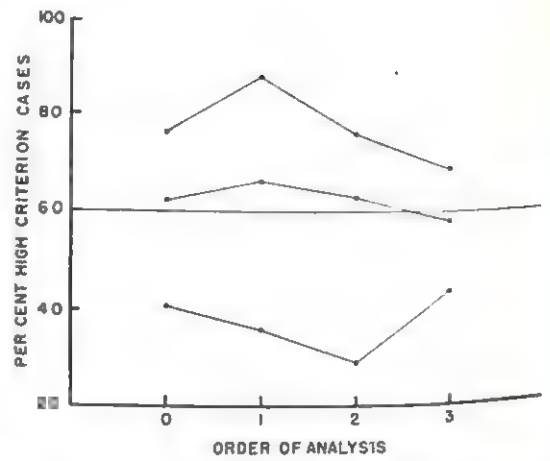


FIG. 1. Shrinkage chart for the Quartile Difference Method of item selection.

items, yields the highest  $Q_1 - Q_4$  value. In this case the third item to be selected was item 1 (Height-Weight Ratio) which, when combined with items 4 and 2, yielded the highest  $Q_1 - Q_4$  value, namely 47.

9. Plot the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values for the best three items as selected by the second order analysis on the shrinkage chart, Figure 1. In this case the  $Q_1$ ,  $Q_{25}$  and  $Q_4$  values plotted for the second order analysis were the values obtained for items 4, 2 and 1. Examination of the shrinkage chart at this point reveals that the addition of item 1 has attenuated the per cent of high criterion cases in  $Q_1$ , but has continued to decrease the per cent of high criterion cases in  $Q_4$ . Since the over-all index of item discrimination, the  $Q_1 - Q_4$  value, shows a drop of five per cent from the first to second order analysis, the researcher might profitably stop selecting items at this point in the selection procedure.

The analysis in this case was continued to include all four items. Computations for this third order analysis are recorded in Table 2 and plotted in Figure 1. Examination of Figure 1 reveals that the inclusion of item 3 results in further shrinkage (the distance between the  $Q_1$  and  $Q_4$  values decreases). In fact, the predictive efficiency of all four items appears to be less than that of the best single item.

The inclusion of the shrinkage chart in the procedure is a refinement but by no means a necessity. The researcher could perhaps as effectively determine when to stop adding items by examining the trend in  $Q_1 - Q_4$  values for each successive composite of selected items as is indicated on the recording sheet, Table 2. Trends in the data for all of the quartile values, however, become more apparent with the shrinkage chart, and for this reason, it is probably well worth the additional labor needed for its construction.

Table 2

Recording Sheet for Quartile Difference Method of Item Selection Computations

Order of Analysis	Item(s)	% High Criterion				Item(s) Selected
		$Q_1$	$Q_{25}$	$Q_4$	$Q_1 - Q_4$ *	
0	4	76	62	41	35	4
1	4,1	73	52	58	15	2
	4,2	88	66	36	52	
	4,3	61	58	37	24	
2	42,1	76	63	29	47	1
	42,3	66	63	32	34	
3	421,3	69	58	44	25	3

\* The highest  $Q_1 - Q_4$  value for each order of analysis is indicated by italics.

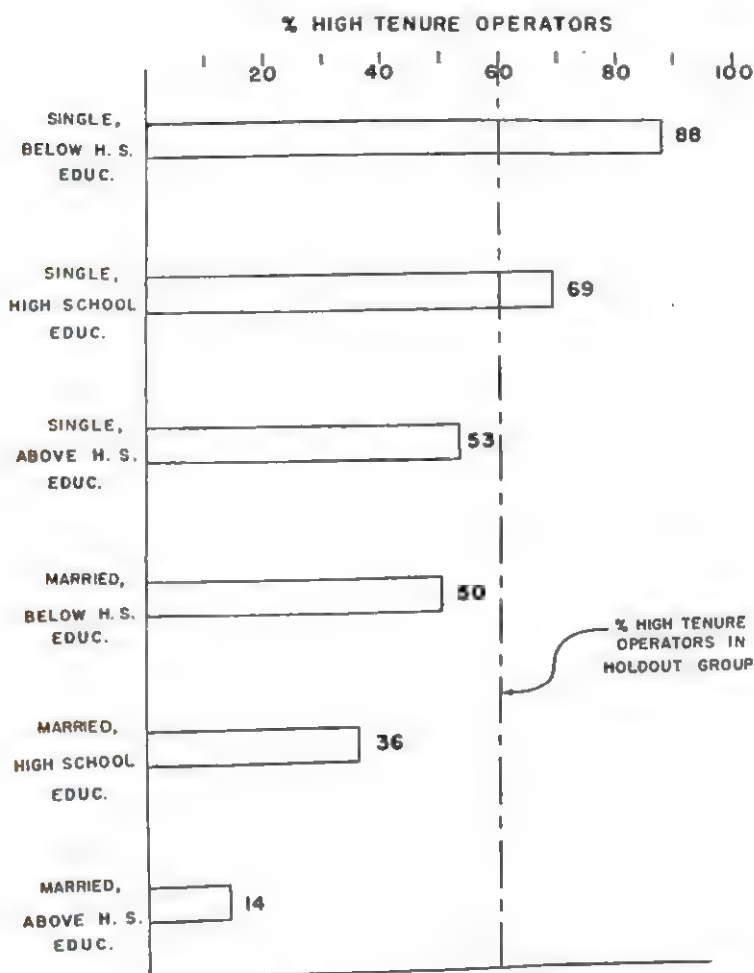


FIG. 2. Per cent of high tenure operators for various combinations of the combined marital status and education categories.

### Results

On the basis of the results provided by the item selection technique, two of the original four items that were considered to be significantly related to tenure (Education and Marital Status) were chosen to compose the selection battery. The per cent of high tenure operators in the holdout group for each combination of response categories for the two items are presented in Figure 2.

It was felt that presenting these results in terms of combined response categories would be more meaningful than presenting them in terms of composite scoring weights. Actually, the listing of combined category responses corresponds to scoring weight magnitudes from

high (top) to low (bottom). A rather impressive and uniform drop in per cent of high tenure operators occurs from the first category (single, below high school education) to the last category (married, above high school education). The trend, while in the right direction, is somewhat stabilized for the two middle categories (single, above high school education and married, below high school education). The per cent of high criterion operators in the former is 53 and in the latter, 50. Chi square was computed for the contingency table composed of frequencies of high and low tenure operators at the six combinations of response categories in order to test the hypothesis that the relationship expressed here could be attributed to chance. The null hy-

pothesis was rejected at better than the 1% significance level.

#### Summary

An item selection technique for categorical data, the quartile difference method, was developed to help the researcher select the most highly predictive combination of items from a pool of possible predictors. The technique while not completely precise (quarter splits have to be approximated and consequently affect the precision of the quartile values for the various analyses) does, however, provide a systematic procedure for the selection of

categorical predictors. The mechanics of the method were demonstrated with four items that were found to be related to the tenure of telephone operators on the basis of item analyses with a primary group. It was found that a combination of two of these items (Education and Marital Status) appeared to be more highly predictive of the criterion than was any other combination of items. In this regard, the per cent of high tenure operators decreases as marital status changes from single to married and as education increases.

*Received December 19, 1952.*

## The Construction of a Personality Scale to Predict Scholastic Achievement<sup>1,2</sup>

Harrison G. Gough

Department of Psychology and Institute of Personality Assessment and Research,  
University of California, Berkeley

This report describes an attempt to develop a brief personality scale to predict college undergraduate course grades, and particularly undergraduate course grades in psychology. The study was undertaken with the expectation that its findings would contribute to a broader understanding of some of the non-intellective factors relating to academic achievement, particularly those factors having to do with personal values, beliefs, and self-definitions. The construction of the scale represents one of a series of studies devoted to the measurement of positive and favorable aspects of personality and individual functioning being carried out by the writer. The present scale, along with a number of earlier scales for such factors as social participativeness, dominance and leadership, social responsibility, and intellectual efficiency, is included as a sub-test in the *California Psychological Inventory*.<sup>3</sup>

The first step in constructing the present scale was to assemble a pool of criterion-specific personality inventory items. The writing and selection of beginning items was based upon three general sources: previous findings, theories about academic motivation and achievement, and intuitive hunches about contributory factors. There is not space in this report to do more than refer to the procedures used in writing and selecting items, but it should be emphasized that a major factor in the possible success of any endeavor such as the present one is the veridicality and au-

thenticity of the items themselves. No amount of analytical precision at some later time can overcome the limitations of an inept, superficial, or tangential pool of items. It is the writer's belief that many psychological studies on the prediction of complex criteria from personality inventory data have floundered because of failure to observe this simple, but fundamental, prerequisite.

Four original samples were obtained for item analysis. These consisted of introductory psychology classes at the University of California, the University of Minnesota, and Vanderbilt University.<sup>4</sup> Each item was studied in at least three of the four samples, and all items revealing discriminatory power in each instance were retained. Table 1 lists five of the items and the basic item analysis statistics.<sup>5</sup>

### The Items

Altogether, 36 items<sup>6</sup> from the pool of 150 items were retained for the first version of the scale, called Hr (for honor point ratio) to distinguish it from an Ac—high school academic achievement—scale developed earlier by the

<sup>4</sup> These samples were very kindly made available by Drs. John Gustad, Rheem Jarrett, and Miles A. Tinker.

<sup>5</sup> A longer version of Table 1 giving the item percentages and significance tests for the complete scale has been deposited with the American Documentation Institute. Order Document 5947 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting \$1.25 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.25 for photoprint readable without optical aid.

<sup>6</sup> Sixteen of these 36 items were taken, by permission, from the Minnesota Multiphasic Personality Inventory. (Hathaway, S. R., and McKinley, J. C. *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press, 1943.) The 16 items in the MMPI and the scored responses are as follows: 33F, 78T, 122T, 157F, 248F, 250F, 260F, 287F, 295T, 313F, 395F, 437F, 448F, 469F, 492F, 498F.

<sup>1</sup> This project was carried out under a research grant from the National Institute of Mental Health, National Institutes of Health, U. S. Public Health Service.

<sup>2</sup> This paper is a revision and extension of a preliminary version given at the annual meetings of the American Psychological Association in Washington, D. C., September 1, 1952.

<sup>3</sup> The complete bibliography for the inventory is too long for inclusion here. For selected references see: (1), (2), (4), (5), and (6).

Table 1

Sample Items from the Hr (Honor Point Ratio) Scale Distinguishing between Students with Higher and Lower Course Grades

Item	Proportion in Each Sample Saying "True"					
	California Class		Minnesota Class		Vanderbilt Class	
	Higher (N=50)	Lower (N=50)	Higher (N=40)	Lower (N=40)	Higher (N=20)	Lower (N=20)
1. Lawbreakers are almost always caught and punished.	44	62	48	58	25	50
2. For most questions there is just one right answer, once a person is able to get all the facts.	22	46	30	38	20	50
3. It is annoying to listen to a lecturer who cannot seem to make up his mind as to what he really believes.	72	92	65	88	65	90
4. The future is too uncertain for a person to make serious plans.	8	22	8	25	35	45
5. Teachers often expect too much work from the students.	32	46	32	52	45	60

writer (4). The 36 items, and the responses predictive of higher grades are given below:

1. I have had very peculiar and strange experiences. (F). 2. I have very few fears compared to my friends. (F). 3. I usually take an active part in the entertainment at parties. (F). 4. It is always a good thing to be frank. (F). 5. I don't blame anyone for trying to grab all he can get in this world. (F). 6. I was a slow learner in school. (F).

7. Sometimes without any reason or even when things are going wrong I feel excitedly happy, "on top of the world." (F). 8. Parents are much too easy on their children nowadays. (F). 9. Teachers often expect too much work from the students. (F). 10. I think I would like to fight in a boxing match sometime. (F). 11. I have often found people jealous of my good ideas, just because they had not thought of them first. (F). 12. People pretend to care more about one another than they really do. (F).

13. The future is too uncertain for a person to make serious plans. (F). 14. The man who provides temptation by leaving valuable property unprotected is about as much to blame for its theft as the one who steals it. (F). 15. I dread the thought of an earthquake. (F). 16. I am bothered by people outside, on streetcars, in stores, etc., watching me. (F). 17. I feel that I have often been punished without cause. (F).

18. I seem to be about as capable and smart as most others around me. (T).

19. I like poetry. (T). 20. It is annoying to listen to a lecturer who cannot seem to make up his mind as to what he really believes. (F). 21. I like to plan a home study schedule and then follow it. (F). 22. Our thinking would be a lot better off if we would just forget about words like "probably," "approximately," and "perhaps." (F). 23. For most questions there is just one right answer, once a person is able to get all the facts. (F). 24. It is all right to get around the law if you don't actually break it. (F).

25. I often lose my temper. (F). 26. I sometimes feel that I am a burden to others. (F). 27. I looked up to my father as an ideal man. (F). 28. Law-breakers are almost always caught and punished. (F). 29. I liked "Alice in Wonderland" by Lewis Carroll. (T). 30. I have a tendency to give up easily when I meet difficult problems. (F).

31. The trouble with many people is that they don't take things seriously enough. (F). 32. Only a fool would try to change our American way of life. (F). 33. Even when I do sit down to study it is hard to keep my mind on the assignment. (F). 34. It is often hard for me to understand what the questions are driving at in a school test. (F). 35. I have to wait for the right mood before I can sit down and study. (F).

Table 2

Summary Statistics for the Original Samples  
on the Hr Scale

Sample	N	M	SD	<i>r</i> with Course Grades
1. Introductory psychology class at California, June, 1950.*	180	15.7	2.6	.42
2. Introductory psychology class at Minnesota, August, 1949.**	67	8.1	2.1	.57
3. Introductory experimental psychology class at Minnesota, October, 1950.	270	24.9	4.0	.47
4. Introductory psychology class at Vanderbilt, October, 1950.	86	21.9	4.4	.47

\* Took only 24 of the 36 items in the full scale.

\*\* Took only 12 of the 36 items in the full scale.

36. I plan very carefully about which school courses I will take. (T).

## Results

This Hr scale was correlated with course grades in the original four samples, totalling 603 cases, with the results indicated in Table 2. The median *r* is .47, and the mean *r*, using the z-transformation, is .48.

Table 3

Summary Statistics for the Cross-Validating Samples  
Given the Full 36-Item Hr Scale

Sample	N	M	SD	<i>r</i> with Course Grades
1. Introductory psychology class at California, March, 1951.	121	23.9	5.7	.35
2. Introductory psychology class at California, June, 1951.	117	22.9	4.2	.26
3. Introductory psychology class at California, August, 1951.	61	23.3	4.0	.31
4. Elementary statistics class at California, October, 1950.	37	24.9	6.2	.58

Table 4

Summary Statistics for the Cross-Validating Samples  
Given the 32-Item Version of the Hr Scale  
Included in the California Psychological Inventory

Sample	N	M	SD	<i>r</i> with Course Grades
1. Introductory psychology class at California, October, 1951.	348	21.3	4.0	.31
2. Introductory psychology class at Stanford, March, 1952.	23	22.5	2.6	.26
3. Introductory psychology class at California, April, 1952.	211	21.0	4.2	.28
4. Introductory psychology class at California, July, 1952.	63	21.3	4.1	.60
5. Introductory psychology class at California (Santa Barbara campus), December, 1952.	104	22.2	3.0	.32
6. Upper division psychology class at California, April, 1952.	139	23.9	3.4	.39
7. Upper division psychology class at California, July, 1952.	29	23.3	3.6	.45

Four cross-validating samples, totalling 336 cases, were given the initial 36-item scale.<sup>7</sup> Table 3 presents the findings. The median *r* here is .33, and the mean *r*, using the z-transformation, is .38.

The original 36-item Hr scale contained four items pertaining to present attendance in school (the last four items in the list above). These items were eliminated in the 32-item version of the Hr scale included in the *California Psychological Inventory*. This inventory was given to seven additional college samples to obtain cross-validated information on the 32-item scale, when included in a large constellation of items.<sup>8</sup>

<sup>7</sup> These samples were made available through the courtesy of Drs. W. Brown, J. McKee, L. Postman, and R. Tryon.

<sup>8</sup> These samples were made available through the courtesy of Drs. W. Brown, J. Clark, P. Farnsworth, D. Krech, D. MacKinnon, and D. Riley.

Table 5

Correlation of the 32-Item California Psychological Inventory Version of the Hr Scale with High School Grade Average

High School	N	M	SD	r
1. Butler, Pennsylvania	397	15.6	4.2	.38
2. Clarksdale, Mississippi	77	14.3	3.9	.26
3. Franklin, Pennsylvania	108	15.8	4.1	.38
4. Mt. Vernon, Washington	107	16.5	4.1	.35
5. Rock Island, Illinois	224	15.1	4.3	.42
6. St. Cloud, Minnesota	195	14.7	3.9	.39

Table 4 presents these data. The total number of cases is 917, and the median  $r = .32$ . The mean  $r$ , using the  $z$ -transformation, is again .38.

Because the *California Psychological Inventory* is designed to be used in high school as well as in college settings, the efficacy of the Hr scale in predicting high school over-all grade averages was determined. Table 5 presents these data.<sup>9</sup> The total  $N$  is 1,108, the median  $r = .38$ , and the mean  $r = .36$ .

The Hr scale, along with a wide variety of other tests, was also given to a sample of 40 senior medical students seen at the University of California Institute of Personality Assessment and Research in an intensive assessment program.<sup>10</sup> Some of the more prominent findings are presented in Table 6.

Perhaps the most important observation here is that the Hr scale correlates with criterion ratings of achievement in medicine as well as it does with undergraduate course grades in psychology. Furthermore, its pattern of correlation with the other variables listed is uniformly favorable, with the possible exception of the staff rating on impulsivity.

One of the questions which might now be raised is whether the Hr scale is assessing any independent achievement variance, or whether it is primarily an indirect measure of intellect. Table 7 affords evidence relevant to this query.

<sup>9</sup> These samples were made available through the kindness of Mr. C. O. Austin, Mrs. M. S. Gleason, Mr. G. N. Harriger, Mr. H. B. Heidelberg, Mr. R. H. Sorenson, and Mr. R. F. Wilson.

<sup>10</sup> The research at the Institute of Personality Assessment and Research is being conducted under a grant from the Rockefeller Foundation. See reference (3) for a discussion of the work of this Institute.

The six correlations with IQ in the high school samples are all lower than they are for Hr vs. grade averages, and a similar difference obtains for the college sample. In the military sample of 150 cases Hr correlates only .10 with intellect, but .50 with a measure of scholastic achievement. The mean  $r$  with the intellectual variables in Table 7 is .26, and with the indices of achievement is .38.

If these values are taken as reasonable approximations of the true parameter values, an estimate of the multiple  $R$  between IQ, Hr, and scholastic achievement can easily be made. For the typical value of .50 between IQ and grades, the multiple  $R$  would be .57, for a value of .60 the multiple  $R$  would be .64, and so on. Hr would thus appear to be a partial

Table 6

Correlation of the Hr Scale with a Variety of Measures and Assessment Variables in a Sample of 40 University of California Senior Medical Students

Variables	r
1. Medical faculty criterion ratings.	
a. Potential success	.31
b. Originality	.33
2. Assessment staff ratings.	
a. Personal tempo	.56
b. Breadth of interests	.51
c. Vitality	.46
d. Impulsivity	.42
e. Verbal fluency	.41
f. Originality	.39
g. Positive affect	.32
h. Rigidity	-.32
3. Ratings of performance in improvisations.	
a. Dominance	.44
b. Flexibility	.34
c. Ingenuity	.33
4. Ratings of performance in charades.	
a. Motility	.42
b. Over-all effectiveness	.34
c. Perseveration	-.27
d. Self-consciousness	-.25
5. Perceptual-cognitive variables.	
a. Size constancy estimation (near and far triangles, smallness of error in judging)	.35
b. Luminous tilted square, total error in adjusting inner line to upright	-.31
c. Street Gestalt pictures, accuracy of recognition	.27

Table 7

Comparative Correlations between Hr and the Intellectual and Achievement Variables Indicated

Sample	N	Correlation of Hr with	
		Intellectual Variable*	Achievement Variable**
I. High Schools			
1. Butler, Pennsylvania	397	.33	.38
2. Clarksdale, Mississippi	77	.10	.26
3. Franklin, Pennsylvania	108	.37	.38
4. Mt. Vernon, Washington	107	.30	.35
5. Rock Island, Illinois	224	.32	.42
6. St. Cloud, Minnesota	195	.33	.39
II. College			
1. University of California, Santa Barbara, psychology class	104	.22	.32
III. Other			
1. Military Officers	150	.10	.50

\* In the high school samples, standard group tests of intelligence were used. In the college sample the criterion was the Altus Measure of Verbal Aptitude, and in the military sample the Thurstone Primary Mental Abilities Test.

\*\* In the high school samples the criterion was the over-all high school grade average, in the college sample the course grade in psychology, and in the military sample the USAFI Test of General Educational Development, Reading Comprehension in the Social Sciences.

Table 8

Correlation of the Hr Scale with Other Scales from the California Psychological Inventory, in a Nationwide High School Sample\*

CPI Scale	Females	Males
1. Re (responsibility)	.55	.46
2. To (tolerance)	.76	.70
3. Fl (flexibility)	.31	.38
4. St (status)	.53	.48
5. Do (dominance)	.31	.24
6. Sp (social participation)	.38	.19
7. Fe (femininity)	-.11	.04
8. De (delinquency)	-.27	-.27
9. Ie (intellectual efficiency)	.68	.55
10. Ac (academic achievement, high school)	.52	.50
11. Py (psychological interests)	.45	.46
12. Ip (academic motivation, graduate school)	.30	.29
13. Ne (neurodermatitis)	-.26	-.26
14. X <sub>1</sub> (poise and spontaneity)	.33	.27
15. X <sub>2</sub> (impulsivity and self-centeredness)	-.40	-.44
16. In (infrequency)	-.07	-.03
17. Gi (good impression)	.39	.37
18. Ds (dissimulation)	-.52	-.46

\* 2,423 females, 2,077 males, from 16 high schools in 13 states.

predictor of academic outcomes in its own right without drawing to any great extent on intellectual factors, and can also add slightly to the multiple R prediction of grades from measures of intelligence.

The intercorrelations of Hr with the 18 other scales on the CPI are presented in Table 8. The highest relationships are with the scales for tolerance, flexibility, intellectual efficiency, and psychological interests.

The final information presented in this paper has to do with the social psychological implications of higher and lower scores on the Hr scale. In the research program at the Institute of Personality Assessment and Research previously referred to, each staff member filled in a Gough Adjective Check List (3) about each assessee. For some of the analyses these observers' reports were composited into a single "general observer's" report by considering each adjective checked by at least 2 out of 6 senior staff members as being "present," and as being "absent" if checked by only one, or by none.

These composited adjective check lists were used to carry out an analysis of the social stimulus values of the Hr scale. Two sam-

ples of 30 each were drawn by selecting the 10 highest and 10 lowest subjects on the Hr scale from two graduate student samples of 40 each, and from the sample of 40 medical school seniors already mentioned. A study was then made of what observers did, in fact, say *about* the 30 highest ranking students, as compared with what they did, in fact, say about the 30 lowest ranking students. The adjectives showing statistically significant differentiations are listed below:

I. Adjectives checked more frequently about higher-scoring subjects on the Hr scale.

adaptable	determined	persevering
alert	efficient	planful
ambitious	fore-sighted	pleasant
appreciative	honest	rational
capable	industrious	reasonable
clear-thinking	intelligent	realistic
conscientious	interests wide	reliable
cooperative	logical	responsible
dependable	organized	resourceful

II. Adjectives checked more frequently about low-scoring subjects on the Hr scale.

cautious	nervous	sentimental
dissatisfied	preoccupied	shy
dull	rebellious	wary
immature	rigid	

The patterning of these adjectives is very consistent. "Highs" are seen as alert, clear-thinking, efficient, intelligent, pleasant, and resourceful. "Lows" are seen as dull, immature, rebellious, rigid, and wary. The staff raters, of course, had no information whatsoever about the Hr scores of these subjects.

### Summary

A personality scale to predict undergraduate grades was developed. A mean  $r$  with course grades of .38 in eleven cross-validating college samples totalling 1,253 cases was attained. The Hr scale also predicted high

school grades, giving a mean  $r$  of .36 in six high school samples totalling 1,108 cases.

Evidence from eight samples, including 1,362 cases, was adduced to support the claim that the Hr scale is a predictor of academic achievement and not simply an indirect and inefficient measure of intellect. In these samples the mean correlation of Hr with measures of intellect was .26, and with indices of academic achievement was .38.

Additional findings in a sample of 40 senior medical students revealed a significant correlation between Hr and ratings of success in medical training, and between Hr and a number of assessment variables such as breadth of interests, originality, flexibility, vitality, effectiveness in group discussion and in charades, and adequacy of performance on perceptual-cognitive tasks involving complex judgmental decisions.

The final section of the paper listed some of the more prominent social-interactional implications of higher and lower scores on the Hr scale. High scorers tend to be seen as capable, intelligent, and reliable and low scorers as dissatisfied, dull, rigid, and shy.

Received November 20, 1952.

### References

1. Gough, H. G. Some common misconceptions about neuroticism. *Psychol. serv. cent. J.*, in press.
2. Gough, H. G. Identifying psychological femininity. *Educ. psychol. Measmt.*, 1952, 12, 427-439.
3. Gough, H. G. *Predicting success in graduate training: A progress report*. Berkeley, California: The University of California Institute of Personality Assessment and Research, 1950. Pp. 1-65 (mimeographed).
4. Gough, H. G. What determines the academic achievement of high school students? *J. educ. Res.*, 1953, 45, 321-331.
5. Gough, H. G., McClosky, H., and Meehl, P. E. A personality scale for social responsibility. *J. abn. soc. Psychol.*, 1952, 47, 73-80.
6. Gough, H. G., and Peterson, D. R. The identification and measurement of predispositional factors in crime and delinquency. *J. consult. Psychol.*, 1952, 16, 207-212.

# Kuder Interest Patterns of Medical, Law, and Business School Alumni

Robert H. Shaffer

*Office of Dean of Students, Indiana University*

and

G. Frederic Kuder

*Duke University*

This paper reports the comparative scores made on the Kuder Preference Record by samples of graduates of the Indiana University Schools of Medicine, Law and Business who had graduated in 1941 or previously. In 1950, the Preference Record (Form C) was sent to 996 graduates of the School of Business, 764 graduates of the School of Law and 992 graduates of the School of Medicine. Returns were received from 313 for Business, 210 for Law and 242 for Medicine. The mean ages of the respondents were 37.5 years for Business, 45.2 for Law and 45.2 for Medicine. Each return indicated the individual's present occupation. Striking and significant differences have been found among the interest patterns of the various groups studied.

The first comparison presented is between the interests of doctors and those of lawyers, accountants, and other Business School graduates. Table 1 gives these data. Since accountants differ so markedly from other business groups they have been kept separate in the table. It should be noted, too, that only

practicing lawyers from the law school are reported in this table. The data from non-lawyers are presented later.

In this and the following tables the mean raw scores reported in the first line are taken as the basis for comparison. The *t*-test was used to determine the significance of differences of means from the base group.

The standard deviations of all groups studied are reported in Table 3.

## Results

Inspection of Table 1 reveals that doctors had scores significantly different at the 1% level from lawyers on one of the ten scales, and from the business groups on seven and eight of the scales. As a general pattern, when compared to the other groups, doctors were higher at the 1% level of significance on the scientific, social service, artistic and outdoor scales and lower at the same level of significance on the computational, persuasive, and clerical scales.

Table 1  
Mean Interest Scores of Lawyers and Businessmen Compared with Those of Doctors

	Out- door	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Serv.	Cleri- cal
Med. Sch. Grads (N = 242)	47.7	40.9	23.9	49.6	30.7	24.5	21.8	12.8	45.1	37.5
Practicing Lawyers (N = 148)	40.1†	33.2†	27.8**	36.6†	41.6**	20.2†	26.5**	13.7	39.2**	50.8**
Accountants (N = 44)	38.0†	39.8	42.1**	35.8†	39.6**	17.5†	23.0	12.0	32.7†	60.3**
Bus. Grads excl. Accts. (N = 269)	37.2†	37.7†	29.8**	34.0†	51.1**	19.6†	21.8	14.1*	39.2†	50.3**

\*\* Significantly higher at the 1% level of confidence.

\* Significantly higher at the 5% level of confidence.

† Significantly lower at the 1% level of confidence.

Table 2  
Mean Interest Scores of Accountants and Other Business School Graduates  
Compared with Those of Lawyers

	Out- door	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Serv.	Cleri- cal
Practicing Lawyers (N = 148)	40.1	33.2	27.8	36.6	41.6	20.2	26.5	13.7	39.2	50.8
Accountants (N = 44)	38.0	39.8**	42.1**	35.8	39.6	17.5	23.0†	12.0	32.7†	60.3**
Bus. Grads excl. Accts. (N = 269)	37.2†	37.7**	29.8*	34.0†	51.1**	19.6	21.8†	14.1	39.2	50.3

\*\* Significantly higher at the 1% level of confidence.

\* Significantly higher at the 5% level of confidence.

† Significantly lower at the 1% level of confidence.

‡ Significantly lower at the 5% level of confidence.

Table 2 gives the data resulting from a comparison of the interest scores of lawyers to the two business groups. As a general pattern the lawyers were significantly higher than businessmen other than accountants in the literary and scientific areas and lower in the persuasive and mechanical areas. The comparison with accountants differed from this pattern. The lawyers had lower computational, clerical and mechanical scores, and higher social service and literary scores than accountants. The comparison of lawyers with physicians was noted in the discussion of Table 1.

It is interesting also to note differences between subdivisions of the graduates of the various schools. Table 3 gives these data. The medical school graduates were scattered among a number of specialties, but some did not report enough detail to allow a more specific classification than that of physician. However, there were enough who could be classified specifically as surgeons and physicians-in-general-practice to justify a comparison of the scores from the two groups. The most significant difference between these groups is on the social service scale, the physicians-in-general-practice being significantly higher within the 1% level of confidence. They are also higher on the scientific scale but only at the 5% level. A trend well within the 10% level of confidence may also be noted for surgeons to be higher on the mechanical scale.

Although it appears that the graduates of the medical and business schools stay in these fields, this generalization does not hold for

the law graduates. Perhaps the distinction between law and business is the result of the terminology used, since "business" covers a tremendously wide range of activities. A person could change his occupation greatly and still be in the field of business. At any rate, of the law school graduates responding, 29.5% reported they were in occupations other than law. Most of these are in related fields often involving managerial or administration work in business and industry where they presumably have occasion to apply their training in law. Those who are not actually practicing law are significantly and perhaps surprisingly higher on the persuasive scale. This is the only significant difference noted between the two groups, as shown in Table 3.

The graduates of the business school are in a wide variety of occupations and except for the accountants the occupational groups were too small to justify a breakdown analysis. Hence the scores of the accountants are compared with those of the remaining graduates of the business school. The results are reported in Table 3. As might be expected, the accountants are significantly higher in the computational and clerical areas. A negative difference of lesser significance may also be noted on the musical scale. A large proportion of the non-accountant group is in managerial or sales occupations. These results are consistent with those previously reported<sup>1</sup> for senior men in the I. U. School of Business. Senior accounting majors were found

<sup>1</sup> Shaffer, R. H. Kuder interest patterns of university business school seniors. *J. appl. Psychol.*, 1949, 33, 489-493.

Table 3  
Mean Interest Scores of Sub-Groupings of Doctors, Lawyers and Businessmen

Group	Outdoor	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Serv.	Clerical
Med. School										
Graduates	M 47.7	40.9	23.9	49.6	30.7	24.5	21.8	12.8	45.1	37.5
N = 242†	SD 14.3	11.8	8.3	8.4	11.4	8.9	7.1	6.5	12.6	10.4
Surgeons	M 51.4	44.5	22.1	47.9	31.5	24.5	22.7	13.4	41.4	37.2
N = 50	SD 11.8	12.5	8.8	9.1	10.4	8.6	6.7	6.9	13.4	11.4
Physicians-in-Gen'l-Pract.	M 48.6	40.6	24.7	51.2*	31.1	23.9	20.4	11.0	48.3**	37.9
N = 66	SD 14.0	10.8	8.2	7.7	12.3	8.4	6.8	6.2	11.6	10.5
Law School Grads.										
Lawyers	M 40.1	33.2	27.8	36.6	41.6	20.2	26.5	13.7	39.2	50.8
N = 148	SD 14.7	13.1	8.1	10.2	13.7	8.6	7.2	6.6	12.0	13.1
Non-Lawyers	M 42.2	34.4	26.1	34.0	47.2**	20.0	26.2	11.9	38.6	50.6
N = 62	SD 15.6	13.1	8.8	8.4	16.3	8.4	7.6	6.2	12.8	14.2
Business School Grads other than Accts.	M 37.2	37.7	29.8	34.0	51.1	19.6	21.8	14.1	39.2	50.3
N = 269	SD 14.0	12.5	9.9	10.3	15.4	8.5	8.1	6.2	11.9	13.1
Accountants	M 38.0	39.8	42.1**	35.8	39.6†	17.5	23.0	12.0§	32.7†	60.3**
N = 44	SD 14.4	12.8	8.0	10.1	14.3	9.1	7.3	5.7	12.2	11.6

† As indicated, 242 Medical School graduates returned questionnaires. A large number of respondents merely stated they were "doctors" instead of indicating actual type of their practice. To prevent erroneous grouping, the returns of only those who indicated they were in "general practice" or "surgeon" were used for statistical comparisons. The *t*-test for Med. School graduates was confined to comparison of surgeons and physicians-in-general-practice.

\*\* Significantly higher at the 1% level of confidence.

\* Significantly higher at the 5% level of confidence.

† Significantly lower at the 1% level of confidence.

§ Significantly lower at the 5% level of confidence.

to have significant differences in all nine scales of the Kuder Form B when compared to all business seniors.

### Summary

The Kuder Preference Record (Form C) was given to a sampling of the 1941 or prior graduates of the Indiana University Schools of Medicine, Law, and Business. The mean raw scores of these groups and sub-groups were compared with the following results:

1. Significantly different interest patterns were found for doctors, lawyers, and businessmen.

2. In general doctors were higher than the other groups on the social service, scientific, artistic and outdoor scales, and lower on the computational, persuasive, and clerical scales.

3. Lawyers compared to businessmen other than accountants were higher in the literary

and scientific areas and lower in the persuasive and mechanical areas. The comparison with accountants revealed a different general pattern. The lawyers had lower computational, clerical, and mechanical scores, and higher social service and literary interest than accountants.

4. Physicians-in-general-practice were found to be higher on the social service and scientific scales than surgeons. There is also a trend of less statistical significance for surgeons to be higher on the mechanical scale.

5. The graduates of the law school who are not practicing law were found to have a higher average persuasive score than the practicing lawyers.

6. Accountants had higher computational and clerical scores than other business groups, and lower social service and persuasive scores.

Received November 10, 1952.

## Kuder Interest Patterns of Student Nurses

Alma Perry Beaver

University of California, Santa Barbara College

A time-honored means of selecting candidates for a given profession or vocation has been the use of the interest or preference inventory. In the opinion of many investigators, interest patterns are more indicative in such selection than are the data afforded by personality schedules and measures of aptitude now in use. In reporting a comparative study of students preparing for five selected professions, Blum states (2): "It is significant that the greatest differences . . . were in their vocational and non-vocational interest tendencies rather than in personality traits. . . ." Triggs (3), drawing upon her wide experience in counseling individual nurses, makes the observation that of those students who fail or withdraw from the nursing curriculum, the most common finding is deviation in scores on the interest inventory; in no other respect is she so likely to deviate from the usual pattern of scores made by the successful nurse. Using the Kuder Preference Record with a group of nurses and a group of women-in-general, Triggs found that it did an excellent job of differentiation. The writer has attempted a similar study with student nurses and liberal arts college girls with an education major.

### The Present Study

The experimental group consisted of 80 students in Knapp College of Nursing in Santa Barbara, California, ranging in age from 17 to 25 years, all Caucasians with the exception of one Japanese girl. Matched individually for sex, age, percentile on ACE, and race, the control group of 50 girls was selected from liberal arts college students with an education major enrolled in the University of California, Santa Barbara College. Table 1 presents these data. The Kuder Preference Record was administered to the student nurses as a part of the battery of qualifying tests given prior to admission to the Knapp College of Nursing. The educa-

tion majors took the inventory on request, as one of a short battery of tests.

Table 2<sup>1</sup> gives the mean percentile scores for each of the nine scales of the Kuder Preference Record for both the experimental and the control groups. Also given are the sigmas for each scale, the standard deviation of the means, the sigmas of the difference and the critical ratios of the difference. Four of the nine scales yield critical ratios at the .01

Table 1

Matching Variables, Experimental and Control Groups

Variables	Experimental Group N = 80	Control Group N = 50
Age, Mean	18.7	18.8
Age, SD	1.5	1.7
ACE, Mean percentile	34.4	35.1
ACE, SD	22.6	22.7

level of confidence. *Science*, with a mean score of 64.6 for the nurses and a mean score of 46.4 for the education majors, has a *t* value of 8.21. Furthermore, the *Persuasive*, *Literary*, and *Social Service* scales also yield highly significant differences between the groups, the respective *t* values being 4.97, 4.17 and 4.94. Figure 1 presents graphically the means of the two groups for all scales of the Kuder Preference Record. Ranked from highest to lowest CR value, the four scales which are least significant in differentiating the groups are *Musical*, *Artistic*, *Mechanical* and *Clerical*.

A less conventional type of analysis of the Kuder Preference Record was also undertaken. The "most" and "least" choices for each item in the clusters of three in all twelve

<sup>1</sup> Tables 2 and 3 have been deposited with the ADI Auxiliary Publications Project. Order Document No. 4068 from ADI Auxiliary Publications Project, Chief, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting \$2.50 for photocopies (6 × 8 inches) or \$1.75 for microfilm.

SOLID LINE = NURSES DASH LINE = COLLEGE FRESHMEN

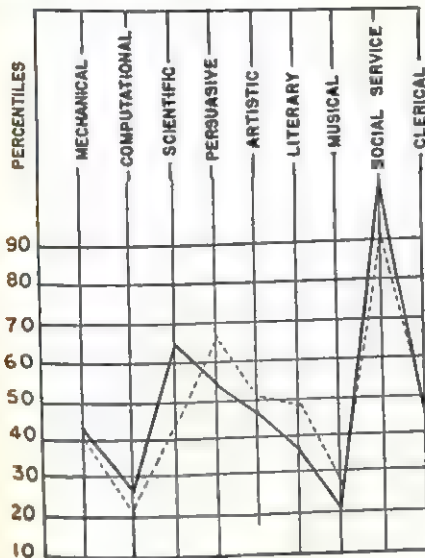


FIG. 1. Comparison of mean percentile scores of 80 cadet nurses and 50 college freshmen on 9 categories, Kuder Preference Test. Significant at the .01 level = science, CR 8.21; persuasion, CR 4.97; literature, CR 4.17; and social service, CR 4.94.

columns of the answer booklet were computed for both the experimental and the control groups. The CR of the percentage of the difference for the items was then computed. A total of 76 items were found to be significant at the .01 level of confidence. This number included 40 items in which the choice of the item as "most" served as the basis for

the differentiation of the groups. Another 16 items were found in which the choice "least" by the groups served as the basis for identification. In still another 20 items, either choice, "most" or "least," yielded  $t$  values at the .01 level of confidence. In sum, then, there were found 76 items which permit 96 choices which appear to be valid for differentiating the student nurse from the liberal arts college education major. The  $t$  values for the 96 choices ranged from 2.62 to 8.22. The highest value was obtained for the item "Be a chemist." It is checked as a "most" choice by the nurses in the cluster that also includes "Be a machinist" and "Be an architect." Table 4 lists samples of the 76 items.<sup>2</sup>

Further examination of these choices in terms of their meaning to the student nurse or education major gives evidence of a rational basis for most of the items. The student nurse is or is expected to be vitally interested in science, chemistry, working in a laboratory and in its equipment, the discovery of cures, and the care of sick people. When an item of this type is checked as a "most" choice by the student nurse, it usually has a high  $t$  value; that is, it distinguishes the student nurse from the education major because members of the latter group seldom check such items as a "most" choice. The educa-

<sup>2</sup> Table 3 in its entirety is deposited with ADI. See footnote 1.

Table 4

Samples of the Significance of the Difference Between the Responses of 80 Student Nurses and 50 College Women in Education Curricula to "Most" and "Least" Choices on the Kuder Record

$t$ Value of Diff.	Item Marking and Group Involved	Item Triads
4.44	"most," Educ.	(r) Take a course in sketching.
5.19	"most," Nurses	(s) Take a course in biology.
		(t) Take a course in metal working.
6.18	"most," Nurses	(R) Do chemical research.
5.77	"least," Educ.	(R) Do chemical research.
4.40	"most," Educ.	(S) Interview applicants for employment.
3.18	"least," Nurses	(T) Write feature stories for a newspaper.
		(G) Write a political campaign song.
		(H) Write an article on how machine tools are made.
		(J) Design a computing machine.
3.53	"most," Nurses	

tion major prefers teaching children, writing a best seller, being a journalist, scoring examinations as a means of earning pin money and interviewing people in a survey of public opinion, to name a few of the interests which her choices reveal. What does the nurse want "least" to do? It will be recalled that the nurse scored low in the *Persuasive* area. She has no interest in selling nor is she interested in writing a newspaper column. Being a journalist or a literary critic or a famous radio commentator has no appeal for her. The college education major, on the other hand, looks with disfavor upon work in chemistry, anything to do with a laboratory or research equipment, and anything associated with a hospital.<sup>3</sup>

A few seemingly odd choices on the part of both groups require interpretation. A "most" choice of the education major, the  $t$  score rating of which is second in magnitude to all the ratings, seems highly peculiar in the light of what the writer believes to be her interests. The item is "Be an architect." This is one of the cluster mentioned previously in which the "most" choice of the student nurse is "Be a chemist." The education major obviously does not want to be a chemist; nor is she interested in being a machinist. She is thus put in the position of making a forced choice, and checks the least offensive item, which is for her "Be an architect." For this same cluster the "least" choice of the education major is "Be a chemist." The critical ratio of the difference for this "least" choice is 3.04. Other possible forced choice answers may be cited. The item "Sell musical instruments" is checked as a "most" item by the education majors. With a  $t$  rating of 5.26 the choice is contrasted with the nurse's choice of "Help in a sick room." Neither wishes to "Repair household appliances." Another example of a forced choice, "Design a computing machine" is checked as a "most" choice by the student nurses in preference to "Write a political campaign song" or "Write an article on how machine tools are made."

In a previous study by the writer (1) of

<sup>3</sup> These students are working toward certification in the fields of early childhood education, elementary education, and physical education at secondary levels.

responses to the MMPI made by student nurses and a matching group of college education majors, 66 items were singled out which differentiated the groups at the .05 level of confidence or better. Of these items, 23 were found to be significant at the .01 level. On the basis of an analysis of these items, presumptive evidence of personality attributes possessed by the student nurse was presented. A point of interest in the present investigation was the possibility of parallel findings or related data in the patterns of response of the student nurse to two inventories designed to explore different facets of personality. Findings are largely negative. Only one striking parallel exists. The two items with the highest  $t$  ratings on the previous study were "I like science" and "I like to read about science" respectively. The two items with highest  $t$  ratings in the "most" choices of the student nurse on the Kuder Preference Record are "Be a chemist" and "Give popular lectures on chemistry." As stated previously, one can readily infer a logical basis for the nurse's choice.

Triggs (3) obtained Preference Record scores on 826 graduate nurses and compared them with the scores of 1246 women-in-general. She found that the interests of nurses differed significantly at the .01 level from women-in-general on all scales of the Preference Record except on the *Artistic*, where the difference was found to be significant at the .05 level only, and on the *Mechanical*, where no significant difference was found. Listed in the order of magnitude, the scales showing positive magnitude were *Social Service*, *Science*, *Artistic* and *Musical*. Those showing negative differences were *Persuasive*, *Clerical*, *Computational* and *Literary* scales. Not closely related to the present study but of interest is another study by Triggs (4) in which she used the scores of the Kuder Preference Record to determine whether different interest patterns exist in specialized fields of nursing. She reports reliable differences; for example, the Public Health nurse makes significantly higher scores on the *Persuasive* and *Social Service* scales, and significantly lower scores on the *Computational* and *Clerical* scales.

### Summary and Conclusions

1. An investigation into the interest patterns of student nurses as contrasted with students majoring in education curricula in a liberal arts college utilized the responses of the respective groups on the Kuder Preference Record. A total of 80 Knapp College of Nursing students were matched for race, age, and percentile on ACE with 50 education majors from the University of California, Santa Barbara College.

2. Mean scores of the two groups in four of the interest areas yielded critical ratios of the difference at the .01 level of confidence. The *Science* scale with a mean score of 64.6 for the nurses and 46.4 for the education majors has the highest *t* value of 8.21. The areas *Persuasive*, *Literary* and *Social Service* yielded respective *t* values of 4.97, 4.17 and 4.94.

3. Another form of analysis was attempted in order to identify items in the clusters of three in which "most" and "least" choices showed a valid difference for the experimental and control groups. A total of 96 choices made in response to 76 items were found to be significant at the .01 level of confidence. Of these, 60 were "most" choices and 36 "least" choices. The *t* values for the individual items varied from a high of 8.22 to a low of 2.62.

4. The student nurse by reason of her "most" choices manifests interest in or preference for science, anything pertaining to the laboratory, the discovery of cures, and the care of the sick. The college education major,

on the other hand, anticipates a liking for teaching children, is interested in various forms of writing, and in interviewing people for public opinion surveys. The "least" choices of the student nurse are heavily weighted with pursuits which require persuasion or selling, any form of writing, reporting or literary criticism. The education major's "least" choices indicate antipathy for work in chemistry, anything that has to do with laboratory or research equipment and anything associated with a hospital.

5. An attempt to find similar personality trends in the response patterns of the student nurse to the Kuder Preference Record and to the MMPI was not successful. A previous study (1) furnished data for the comparison. In one respect only were the findings comparable. The two items in the MMPI receiving the highest *t* value referred to a liking for science. Similarly in the Kuder, the two "most" choices of the nurses having the highest rating referred to interest in being a chemist or giving lectures in chemistry.

Received June 2, 1953.

Early publication.

### References

1. Beaver, Alma P. Personality factors in choice of nursing. *J. appl. Psychol.*, 1953, 37, 374-379.
2. Blum, L. P. A comparative study of students preparing for five selected professions including teaching. *J. exp. Educ.*, 1947, 16, 31-65.
3. Triggs, Frances O. The measured interests of nurses. *J. educ. Res.*, 1947, 41, 25-37.
4. Triggs, Frances O. The measured interests of nurses: a second report. *J. educ. Res.*, 1948, 42, 113-121.

# Personality Factors in Choice of Nursing

Alma Perry Beaver

University of California, Santa Barbara College

Screening devices for choosing candidates for a given profession or occupation have in the main centered in the cognitive rather than the affective areas. Measurement of aptitudes has achieved a degree of objectivity and validity that is still somewhat rare in the less tangible areas of personality and motivation. Yet the importance of the latter factors has not gone unnoticed. In the field of nursing subjective evaluations have pointed to certain intrinsic personality qualifications as essential to success. Disciplined efficiency under emergency conditions and the ability to give comfort and reassurance to the patient are among the demands made upon the nurse. More specific identification of essential traits and a means of measuring them is a goal not yet achieved.

The belief that nurses as a group do represent a more stable segment of the population than the average has had popular acceptance for some time. Studies bulwarking this belief have not been lacking. In 1927 Elwood (1) studied two groups, one made up of nurses, the other college girls. Using Laird's *Introvert-Extrovert Scale* and Woodworth's *Emotional Inventory*, he concluded that both tests placed the nurse in a more favorable light. Lough (3), using the responses on the *Minnesota Multiphasic Personality Inventory* as a basis for her analysis, compares nursing students with women students enrolled in liberal arts and education curricula. She reported the nurses as being more stable than the other groups and as having more masculine interests. In a subsequent study

Lough (4) substantiates her findings through a statistical validation of the differences found between the cadet nurses and the students of General Curriculum. Healy and Borg (2), using the Guilford-Martin battery of personality tests measuring thirteen putative factors, compared a group of nursing-school freshmen from six schools of nursing with students at the University of Texas. They found no characteristic pattern in the analysis of scores of the beginning nurses. They state that this is to be expected to some extent since the students are not screened in most of the schools and the data were collected prior to the withdrawal of students not fitted to the program.

## The Present Study

The writer's study investigating the personality attributes of student nurses is also a comparison between student nurses and a group of college women students majoring in education curricula. A total of 86 women students enrolled in the Knapp College of Nursing at Santa Barbara, California, made up the experimental group. These were matched for race, sex, age and percentile on the ACE, with an equal number of education majors at the University of California, Santa Barbara College. There was one Japanese in each group, the remainder being Caucasian. The age range for both groups was 17 to 25 years. Means for the groups for age and for ACE percentiles are presented in Table 1. Students were individually matched for these variables. In the majority of cases age was

Table 1  
Matching Variables, Experimental and Control Groups

	N	Age		ACE Percentile	
		Mean	SD	Mean	SD
Experimental Group	86	18.7	1.5	34.4	22.6
Control Group	86	18.7	1.7	35.1	22.8

also held constant or varied by not more than two years. Variation in ACE percentile points was not greater than three points except for a very small number of cases.

The group MMPI was administered to all students prior to entering college. The score sheets were then analyzed to determine whether a group of questions could be identified which would differentiate one group of students from the other. A total of 66 items were singled out, the critical ratio of the percentage difference being 2.00 or greater for each of the 66 items (23 of these items were significant at the .01 level). These items were broken down into four named categories and one miscellaneous group by the investigator. These categories, which presumably identify personality characteristics of the student nurse as contrasted with those of education majors, are presented in Table 2. Also given are the item number of the Group MMPI, the  $t$  value of the difference, and the answer characteristic of the nursing student.

The first category of ten items, labeled "A Social-Sexual Factor," is characterized by a preference for the mannish and for masculine activities. The student nurse admits a preference for association with her own sex; she likes the tall mannish woman. There is nothing of the feminine coquette in her make-up nor does she find pleasure in social dancing. Soldiering, reporting sports news, forestry work, all have their appeal for her.

The second group includes twelve items and seems to delineate a conventional adherence to custom and a prudish, decorous attitude usually absent in today's coed. She is embarrassed by dirty stories. She avoids sexy shows; she is not in on the gossip of the group. She disapproves of women smoking and imbibing alcoholic beverages. She does not believe that men are absorbed with the subject of sex. A person should be punished for breaking the law, even if it is unreasonable. Duty to a life goal inspires and motivates her.

Minimal psychosomatic concern is the label given to the third group consisting of eleven items. Two of these items have high  $t$  values. The first one, "The sight of blood neither frightens me nor makes me sick," with a  $t$

value of 4.33, ranks highest in the total list with the exception of two items in the miscellaneous group which refer to a liking for science. A  $t$  value of 3.83 is obtained for the other item, "I sweat very easily even on cool days." The student nurse does not worry about her health, does not dread seeing a doctor, can't remember "playing sick," does not feel tired. Symptoms of hypochondria are lacking.

The responses by the student nurse to the items making up the fourth category, labeled "Freedom from Neuroticism," are indicative of emotional stability. This category contains the largest number of items, twenty in all. The difference is significant at the .01 level of confidence for seven items. The student nurse denies fear of people, the dark, and high places. Anxiety and tension are not part of her everyday experience. Home life is pleasant and few quarrels with members of her family are admitted. She makes no claim to personal importance, but states that she expects to succeed in the activities which she attempts.

The miscellaneous category, including twelve items, presents some difficulties of interpretation. Responses to some items, as the first two, referring to a liking for science and science reading, are easily understood on a common sense basis. As previously stated these two items have the highest  $t$  value of the entire list, being respectively 8.17 and 7.66. It seems fairly obvious that candidates for nursing training would check such items as true. Another group of responses seems almost paradoxical in the light of previous interpretations. Items 102, 461, and 417, all answered as true by the nurse, might be interpreted as contradictory of claims made for freedom from neuroticism. The writer has no rationalization to offer. These items refer respectively to the hardest battles being with oneself, difficulty in setting aside a task once begun, and annoyance with anyone who tries to get ahead in line. Three items, 89, 199 and 261, of rather low discrimination, seem to have little significance for this study as far as interpretation is concerned. Three other items, all answered false, have reference to occupational preference and membership in

Table 2

The Significance of the Difference Between the Responses of 86 Student Nurses and 86 College Women in Education Curricula to 66 Categorized Items on the Group Form of the Minnesota Multiphasic Personality Inventory

MMPI Item Number	t Value of Diff.	Marking Characteristic of Student Nurses	Categories and MMPI Questions
<i>A Social-Sexual Factor Characterized by a Preference for the Mannish and Masculine Activities</i>			
514	3.00	True	I like mannish women.
69	2.60	True	I am very strongly attracted by members of my own sex.
435	2.00	True	Usually I would prefer to work with women.
391	2.00	False	I love to go to dances.
441	2.14	True	I like tall women.
144	2.00	True	I would like to be a soldier.
283	2.00	True	If I were a reporter I would very much like to report sporting news.
561	2.66	True	I very much like horseback riding.
81	2.14	True	I think I would like the kind of work that a forest ranger does.
208	2.00	False	I like to flirt.
<i>A Conventional Attitude</i>			
427	2.71	True	I am embarrassed by dirty stories.
455	2.66	True	I am quite often not in on the talk and gossip of the group I belong to.
378	2.57	True	I do not like to see women smoke.
457	2.40	True	I believe that a person should never taste an alcoholic drink.
232	2.28	True	I have been inspired to a program of life based on duty which I have since carefully followed.
548	2.14	True	I never attend a sexy show if I can avoid it.
183	2.00	True	I am against giving money to beggars.
485	2.00	False	When a man is with a woman he is usually thinking about things related to her sex.
135	2.66	False	If I could get into a movie without paying and be sure I was not seen I would probably do it.
456	3.00	False	A person shouldn't be punished for breaking a law that is unreasonable.
82	2.17	True	I am easily downed in an argument.
99	2.29	False	I like to go to parties and other affairs where there is lots of loud fun.
<i>Minimal Psychosomatic Concern</i>			
128	4.33	True	The sight of blood neither frightens me nor makes me sick.
174	2.29	True	I have never had a fainting spell.
193	2.29	True	I do not have spells of fever or asthma.
55	2.17	True	I am almost never bothered by pains over the chest or in the heart.
36	2.14	True	I seldom worry about my health.
412	2.00	True	I do not dread seeing a doctor about a sickness or injury.
363	3.83	False	I sweat very easily even on cool days.
481	2.00	False	I can remember "playing sick" to get out of doing something.
506	2.00	False	I am a high-strung person.
544	2.60	False	I feel tired a good deal of the time.
155	2.00	False	I am neither gaining nor losing weight.
<i>Freedom from Neuroticism</i>			
137	3.20	True	I believe that my home life is as pleasant as that of most people I know.
96	2.17	True	I have very few quarrels with members of my family.
32	4.00	False	I find it hard to keep my mind on a task or job.
73	3.50	False	I am an important person.

Table 2—Continued

MMPI Item Number	t Value of Diff.	Marking Characteristic of Student Nurses	Categories and MMPI Questions
165	3.14	False	I have several times had a change of heart about my life work.
163	4.00	True	I do not tire quickly.
370	3.00	False	I hate to have to rush when working.
352	2.20	False	I have been afraid of people or things that I knew could not hurt me.
388	2.20	False	I am afraid to be alone in the dark.
166	2.00	False	I am afraid when I look down from a high place.
356	2.00	False	I have more trouble concentrating than other people seem to have.
480	2.00	False	I am often afraid of the dark.
351	2.66	False	I am anxious and upset when I have to make a short trip away from home.
534	2.00	True	Several times I have been the last to give up trying to do a thing.
13	2.20	False	I work under a great deal of tension.
345	2.20	False	I often feel as if things were not real.
271	2.00	True	I do not blame a person for taking advantage of someone who lays himself open to it.
257	2.40	True	I usually expect to succeed in things I do.
426	2.20	True	I have at times had to be rough with people who were rude or annoying.
493	2.43	True	I prefer work which requires close attention to work which allows me to be careless.
<i>Miscellaneous</i>			
221	8.17	True	I like science.
552	7.66	True	I like to read about science.
102	2.29	True	My hardest battles are with myself.
461	3.29	True	I find it hard to set aside a task that I have undertaken, even for a short time.
417	3.20	True	I am often so annoyed when someone tries to get ahead of me in a line of people that I speak to him about it.
89	2.00	True	It takes a lot of argument to convince most people of the truth.
199	2.00	True	Children should be taught all the main facts of sex.
261	2.00	True	If I were an artist I would like to draw flowers.
204	3.83	False	I would like to be a journalist.
428	2.29	False	I like to read newspaper editorials.
387	3.40	False	The only miracles I know of are simply tricks that people play on one another.
229	2.43	False	I should like to belong to several clubs or lodges.

clubs. Similar items answered as true seemed to fit into the Social-Sexual category. Is it because the latter had a definitely masculine connotation for the student nurse or is some other elusive factor operative? Item 387, which refers to miracles as simply tricks played by people on one another, as answered by the student nurse may have reference to some sort of youthful idealism which she entertains regarding possible miracles performed by the members of the medical profession. Divine aid perhaps being assumed. The t

value of 3.29 signifies high validity for this item.

As a means of validation, the 66 items were mimeographed on a single sheet and the students were asked to recheck the list. This presents a slightly different situation from that obtaining in the first instance when the 66 items were scattered through the total aggregate of MMPI items. Furthermore, the cadets were now established in their training program whereas previously they were not certain of admission to the nursing school.

Mean scores were computed for both the nursing and the college groups and compared with the original averages for the 66 items. In each instance, the average number of plus answers was lower than in the original tests, but the difference between the averages of the groups remained approximately the same. A further check was attempted when two groups of nursing students from the Bishop Johnson College of Nursing and the Hollywood Presbyterian Hospital School of Nursing, both in Los Angeles, California, were asked to check the 66 items. The mean score of the latter group, from Hollywood Presbyterian School, was very close to that of the original mean score and higher than the mean retest score of the Knapp group. In the Bishop Johnson College group, the mean score was very close to that of the Knapp retest score. When the significance of the difference between the mean scores of the nurses and the college women was tested, CR's were found to vary from 5.15 to 11.30. The lowest CR was obtained from a comparison of the mean scores of Santa Barbara students and Bishop Johnson School nurses; the largest CR, 11.30, was obtained from a comparison of the mean scores of Santa Barbara students and the Knapp College students on the original test. Additional comparisons yielded scores intermediate between these extremes. These data are given in Table 3. The Pearson product reliability of the 66 item test as measured by the split-half, odd-even technique was .64.

### Summary

1. An investigation into the personality attributes of student nurses as compared with education majors utilized the responses of the group MMPI as the basis for study. A total of 86 women students enrolled in the Knapp College of Nursing at Santa Barbara were matched for race, age and ACE aptitude percentile with an equal number of education majors at the University of California, Santa Barbara College.

2. From the total number of MMPI responses 66 items were singled out to make up a scale which differentiated one group from the other. The criterion of selection was a  $t$  of 2.00 or greater. Of the 66 items, 23 were found to be significant at the .01 level of confidence. The odd-even, split-half technique yielded an  $r$  of .64 as the reliability of the 66 point test.

3. The 66 point test gave mean scores which differentiated the two groups to approximately the same degree as did the original tests, although the absolute scores were lower in each case. The CR of the mean difference between the two groups in the original test was 11.30. When the students were retested with the 66 item test, the CR was 9.98. The difference in mean scores between the two groups in the original test was 9.38 and in the retest 9.68. Two other groups of student nurses in Los Angeles schools who were tested on the 66 item test yielded average scores similar to the experimental groups. The CR's of the mean difference for these respective

Table 3  
Mean Scores for Sixty-six Items, Original and Retest Data

Student Groups	Original Tests				CR
	N	Mean	Sigma	SF <sub>100</sub>	
1. Knapp College of Nursing*	86	41.6	4.5	.48	1 vs. 2
2. Santa Barbara College*	86	32.2	6.2	.67	2 vs. 3
3. Bishop Johnson School of Nursing	50	37.1	4.7	.67	2 vs. 4
4. Hollywood Presbyterian School of Nursing	57	40.0	5.5	.73	5 vs. 6
5. Knapp College of Nursing	54	38.8	Retest		
6. Santa Barbara College	36	29.1	4.4	.75	

\* Sixty-six items embodied in total matrix of MMPI.

groups when compared with the Santa Barbara College group were 5.15 and 7.95.

4. The 66 items were broken down into four categories presumably identifying personality attributes of the student nurse. These were labeled a Social-Sexual Factor, a Conventional Attitude, Minimal Psychosomatic Concern, and Freedom from Neuroticism. An additional group of 12 items made up a miscellaneous category.

5. The study offers evidence that the student nurse presents a significantly different pattern of response for a small number of selected items on the group MMPI when compared with a group of college education majors. Presumptive evidence is furnished that the student nurse is a more stable individual who exhibits a preference for her own sex and likes mannish qualities in her associates. She is fastidious and conventional in her attitude and is duty inspired. Symptoms of hypochondria are lacking as is

evidence of neuroticism. These findings, in general, corroborate the findings of Elwood and Lough. Though both Lough and the writer used the group MMPI as the basis for analysis, the approach is somewhat different. Lough utilized the mean scores from the MMPI profile while the writer used the individual item responses.

Received January 9, 1953.

#### References

1. Elwood, R. H. The role of personality traits in selecting a career; the nurse and the college girl. *J. appl. Psychol.*, 1929, 1, 199-201.
2. Healy, I. and Borg, W. R. Personality characteristics of nursing school students and graduate nurses. *J. appl. Psychol.*, 1951, 35, 265-280.
3. Lough, Orpha M. Women students in liberal arts, nursing and teacher training curricula and the MMPI. *J. appl. Psychol.*, 1947, 31, 437-445.
4. Lough, Orpha M. Correction for "Women students in liberal arts, nursing, and teacher training curricula and the MMPI." *J. appl. Psychol.*, 1951, 35, 125-126.

## Item Validity of the Lee-Thorpe Occupational Interest Inventory

Leopold Bridge and Meyer Morson

Baltimore Regional Office, Veterans Administration<sup>1</sup>

During the use of the *Occupational Interest Inventory*, *Advanced* for the identification of specific interest areas it was noticed that some of the test items did not appear to relate closely to the specific areas for which they were scored. This apparent discrepancy led the authors of the present article to explore the general concepts of validity underlying the construction of the Lee-Thorpe Inventory to determine whether their observations had any bearing on the usefulness of the test.

A search of the available published material showed little relating to the validity of the Occupational Interest Inventory. Super wrote in 1949 (6) that he had located no studies of its validity; and no validity studies are reported in the *Third Mental Measurements Yearbook* (3). A study by McPhail (5) described the establishment of interest profiles for various occupational groups through the use of the Inventory but did not examine the validity of the test as a measuring instrument.

No validity data are presented in the *Manual of Directions* (4) for the Inventory but the authors state that the observation of the following criteria has contributed to the validity of the tests: (1) the selection of the items; (2) the design or description of the items; (3) the balance of the items constituting the Inventory; and (4) the presentation of the items.

The importance of these criteria is apparent from an inspection of the test. This study concerns only Part I which consists of 120 pairs of items each member of a pair being identified by the authors with one of six major occupational fields designated by a letter

and a descriptive phrase: (A) Personal-Social (P-S.); (B) Natural (Nat.); (C) Mechanical (Mech.); (D) Business (Bus.); (E) The Arts (Ar.); and (F) The Sciences (Sci.).

Each major field contains 40 items each presumably descriptive of the respective field. It is, obviously, essential that extreme care be exercised in the selection of each test item when a total of 120 responses will result in determining the order of preference among six areas of interest.

Unless an activity is properly designated, preference for it will distort two scores—the field for which it is scored and the field to which it really belongs. In the event that the items are not properly representative of the occupational fields for which they are scored it would be expected that the testees would express disagreement with their inventoried interest patterns or with the relative rankings of their inventoried interests.

In two studies Brown (1, 2) compared the expressed and inventoried interests of veterans as measured by the Lee-Thorpe Occupational Interest Inventory and found that there were significant differences. Although Brown did not discuss the question it is possible that some of the discrepancy was due to the fact that items in the test were improperly designated as to the occupational field.

### Method and Procedure

In order to evaluate the hypothesis that some of the Lee-Thorpe items were not properly designated it was decided to analyze each item to see if it was assigned to the proper occupational field. It was felt that the persons best qualified to make such a determination would be those who had considerable experience in occupations either from the point of view of counseling or occupational analysis. The items were therefore reviewed by a group of 38 raters, including 18 Counselors or Occupational Analysts from the Maryland State Employment Service with an average experi-

<sup>1</sup> This work was not performed in connection with Veterans Administration activities and does not involve VA records. The opinions expressed are those of the authors and not necessarily those of the Veterans Administration. The authors wish to express their sincere appreciation for the cooperation of Miss Banos of the Maryland Employment Service, Dr. Spool of the Veterans Administration, and Dr. Terwilliger of the Maryland State Vocational Rehabilitation Division, and their respective staffs.

Table 1  
Analysis of Dissent Scores for Major Occupational Fields

Field of Interest	No. of Items	Items Questioned	Items Not Questioned	No. of Dissents	Mean Dissents Per Item
P.-S.	40	14	26	165	4
Nat.	40	26	14	341	9
Mech.	40	30	10	596	15
Bus.	40	12	28	76	2
Ar.	40	23	17	305	8
Sci.	40	28	12	346	9
All Fields	240	133	107	1829	8

ence of 12 years, 15 Vocational Counselors from the Maryland State Rehabilitation Division whose experience averaged  $4\frac{1}{2}$  years, and 5 Vocational Advisers from the Veterans Administration with an average counseling experience of 6 years.

Each person was asked to assign each of the 240 test items to the Field of Interest in which he felt that it belonged. The possibility of being influenced by the prior designations of the test constructors was eliminated by securely masking the letter designations of the items printed in Part I of the Inventory.

The responses of the 38 qualified raters were scored in terms of *dissents*, i.e., assignment of an item to a Field other than that designated by the authors. The few instances where raters felt that an item did not fit into any of the six fields of interest were also scored as *dissents*. On this basis the *dissent score* for each of the 240 items could have any value from 0 to 38. A score of 0 indicates that all the raters assigned the item to the same field as the authors of the test while a score of 38 indicated that none of the raters felt that the item belonged in the field to which it had originally been assigned.

### Results

After tabulating the dissent scores by item and by occupational field a wide scattering of scores was apparent. Opinion ranged from

1. No Dissents ..... Raters place item in the same occupational field as the authors.
2. Low Dissent Score ..... Some disagreement as to the proper placement of the item but too slight to justify positive assertion that item is improperly placed.

total agreement to total disagreement with a general tendency to cluster at the extremes. It was found that some degree of dissent was found for 133 of the 240 items. The Fields which had the least number of questioned items were the Business and Personal-Social while the greatest number of challenged items were in the Mechanical and Scientific Fields.

In view of the fact that *dissent scores* ranged from 0 to 38, it was felt that the intensity of the dissent per item should also be considered in evaluating the soundness of the various fields of interest.

Table 1 indicates that the Business Field was considered as the soundest area in the test. In addition to having the fewest items questioned, it also has the lowest total number of dissents. On this same basis the six fields of the Occupational Interest Inventory may be ranked from most valid to least valid as follows: 1. Business; 2. Personal-Social; 3. Arts; 4. Natural; 5. Scientific; and 6. Mechanical. This order expresses the relative degree to which the items in the test were considered to actually correspond to the field for which they are scored and which they purport to measure.

Further analysis of the dissent scores can be used for an evaluation of the individual item validities. On the basis of these scores the test items can be divided into three major categories:

3. Significant Dissent .....Raters believe that item does not belong in field assigned by authors.
- A. High Agreement among raters ..When there is a high degree of agreement among the dissenting raters the item may be considered as sound but belonging in a field other than that assigned by the authors.
- B. Disagreement among raters ....Item is not sound because there is no general agreement as to the field to which it should be assigned.

As has been previously stated there are 107 items (out of 240) concerning which the raters are in complete accord with the authors of the test. Further, if we assume that a dissent score of 12 or less (at least 66% agreement with the authors) indicates a satisfactorily classified item, we find that 76 additional items or a total of 183 should be considered as assigned to the proper occupational field. In other words, the raters agree that 76% of the questions in Part I of the Occupational Inventory meet the necessary criteria of validity for inclusion in the test.

There are 24 items with dissent scores of 26 or more which indicates substantial disagreement as to the authors' classification of the items but where, in addition, there are significant agreements among the raters as to the proper placements of the items. Of these items 11 are now classified in the Mechanical Field, 4 each in Arts and Sciences, 3 in Natural and 2 in Personal-Social. As a result of the raters' revisions, 5 should be in the Personal-Social Field, 3 in the Natural, 3 in the Mechanical, 3 in the Business, 4 in the Arts, and 6 in the Sciences. The following examples are characteristic of the items that the raters scored as being in improper categories:

Occupational Field Assigned by		
Authors	Raters	Item as Shown in Inventory
Nat.	Sci.	"Plan experiments to control worms, insects and other pests."
Ar.	P.-S.	"Teach people how to improve their manners and poise."
Sci.	Mech.	"Clean and recharge storage batteries."
Mech.	Art.	"Paint signs on windows or do lettering on posters with brush or pen."
Mech.	Sci.	"Experiment with the making of synthetic products, such as artificial teeth, nylon or cellophane."

Up to this point the discussion has accounted for 207 items. The balance consists of the 33 controversial items where no substantial number of the raters could agree as to the proper occupational designation. The following items are characteristic of this group (the numbers in parentheses indicate the number of raters preferring the occupational field designated):

Occupational Field Assigned by		Item as Shown in Inventory
Authors	Raters	
P.-S.	P.-S. (23) Bus. (15)	"Take care of the correspondence and private affairs of another person."
Nat.	Nat. ( 9) Sci. ( 9) Bus. (19) Mech. ( 1)	"Direct the quick-freezing or dehydration of farm products."
Mech.	Mech. ( 5) Bus. (22) P.-S. ( 5) Sci. ( 1) Nat. ( 5)	"Label bottles, sort and wrap fruit, or pack eggs."
Art.	Art. ( 3) P.-S. (12) Nat. (23)	"Mow lawns, clip hedges and bushes, trim trees."
Sci.	Sci. (14) P.-S. (20) Mech. ( 4)	"Keep a doctor's tools and equipment in order."

A careful inspection of these items will indicate the basis on which the raters based their opinion. While there is some defense for each choice there is no way to show that the item will not be regarded in various ways by persons taking the test and therefore cannot be said to meet the criteria of validity as set forth in the *Manual*.

A study of the results of this survey compares in an interesting fashion with the study made by Brown with 60 veterans which involved a comparison of expressed and inventoried interests as shown by the *Lee-Thorpe Occupational Interest Inventory* (2). One of the conclusions reached was that 74.4 per cent

of the veterans felt that their expressed interests corresponded with the relative ranking of their interests as shown by the test. He found that the greatest dissent was found in connection with scores in the Mechanical Field with a bias toward the belief that the scores were too low. A review of the test items by a group of raters experienced in the fields of vocational counseling, placement and job analysis reveals that 11 of the 40 items now classified as "Mechanical" actually belong in other fields but that only three items otherwise classified should be considered as "Mechanical." This leaves the test with only 32 items in this field so that in many instances mechanical interests may be inadequately measured and may account for the expressed dissatisfaction as found by Brown.

The results of the present study indicate that room exists for further detailed work on the selection of items for inclusion in the Inventory. The problem of selecting items that can be considered as belonging exclusively to one field is difficult but not impossible as shown by the number of items on which the raters were in complete accord with the authors. It is essential, however, to see that the items do not contain activities or elements that belong to more than one field. It would be even more important to demonstrate that the items also do, in fact, discriminate between occupational groups a la Strong.

#### Summary and Conclusions

1. The 240 items of Part I of the *Lee-Thorpe Occupational Interest Inventory, Advanced* were analyzed in terms of agreement of raters with the occupational fields for which they are scored.

2. The analysis was performed by 38 raters with extensive backgrounds in vocational counseling and occupational analysis.

3. Scoring was done in terms of dissents, i.e., disagreement with the authors' classification of the items.

4. Raters were in complete accord with the authors of the Inventory on 107 items and in substantial agreement on 76 more items.

5. Raters felt that 24 items were in occupational fields other than those in which they are now assigned, the Mechanical Field being least reliable with 11 out of 40 items considered to be improperly classified.

6. No substantial agreement was reached as to the proper occupational classification of 33 more items.

7. Since the validity of 57 of the 240 items is questionable, caution should be used in the interpretation of the interest pattern obtained through use of the Inventory.

Received December 10, 1952.

#### References

1. Brown, M. N. Expressed and inventoried interests of veterans. *J. appl. Psychol.*, 1951, 35, 401-402.
2. Brown, M. N. Evaluation of Lee-Thorpe inventory ratings by veteran patients. *Educ. psychol. Measmt.*, 1951, 11, 248-254.
3. Buros, O. K. *Third mental measurements year-book*. New Brunswick, Rutgers University Press, 1949.
4. Lee, E. A. and Thorpe, L. P. *Manual of directions, Occupational Interest Inventory, advanced series*. Los Angeles, California Test Bureau. Copyright 1943 (but containing references to test changes made in 1946).
5. McPhail, A. H. Interest patterns for certain occupational groups: Occupational Interest Inventory (Lee-Thorpe). *Educ. psychol. Measmt.*, 1952, 12, 79-89.
6. Super, D. E. *Appraising vocational fitness*. New York: Harpers, 1949.

## Scalability and Validity of the Socio-economic Status Items of the Purdue Opinion Panel

H. H. Remmers and R. Bruce Kirk

Purdue University

In each of the polls of the *Purdue Opinion Panel* a brief scale is included to measure the socio-economic status of the respondents in a nationally representative sample of high school youth numbering from 8,000 to 18,000. The purposes, scope and details of the operation of *Panel* have been described elsewhere (1, 2, 5, 6). The items in this brief scale were originally taken from *The American Home Scale* as the items with highest validity (4). They have been slightly revised on occasion because of obsolescence of an item. For example, an item asking whether any member of the family had been on relief was valid in 1940, but obviously not in 1953. The items used in the present study are as follows.

House and Home: Answer these questions by checking "yes" or "no" in the space below.

"Does your family have:

- |   |                   |
|---|-------------------|
| A. a vacuum cleaner?  | Yes..... No.....  |
| B. an electric or gas refrigerator?   | Yes..... No.....  |
| C. a bathtub or a shower with running water?  | Yes..... No.....  |
| D. a telephone?   | Yes..... No.....  |
| E. an automobile?   | Yes..... No.....  |
| F. Have you had paid lessons in dancing, dramatics, expression, elocution, art, or music outside of school? | Yes..... No....." |

### The Problem

The problems of the present study were: (1) testing the unidimensionality of these items by means of the Guttman test of scalability (3); and (2) testing the validity of the scale.

### Procedure

Two independent random samples of 100 respondents' records were drawn from a total available sample of about 10,000 by taking

every  $n$ th individual record. Each sample of 100 was then tested by means of the Guttman technique (3), once by restricting cut-off points to score boundaries, and again by the "ideal" method.

All pupils in the two samples reported at least one of the items. Possession of an automobile did not, however, distinguish anywhere along the line, thus showing it to be a non-discriminating item. Quite possibly the make, year and model of the automobile might be found to be relevant to such an index of socio-economic status, but this information was not in hand. The other items scale satisfactorily by Guttman's criterion of 90 per cent reproducibility.

### Validity of the Scale

Validity may be variously defined but perhaps its most acceptable meaning is in terms of the prediction of a criterion. It is essentially in this sense that we use the term here. The data are taken from *Purdue Opinion Panel* Poll Reports in the form of statistically reliable differences between the low and the high status groups as defined by the scale. All differences reported here are at the 1 per cent level of confidence or better. The critical ratios range from 2.6 to 9.5. The stratified-random sample is usually between 2,000 and 3,000 respondents with from 20 to 25 per cent in the high socio-economic status group and from 75 to 80 per cent in the low group.

In Poll No. 21 in 1949<sup>1</sup> the students in the national sample were asked to check their individual problems in a list of 300 such problem items. The following items in Table I summarize the breakdown of responses which yielded reliable differences, i.e., reliable cor-

<sup>1</sup> Results of this poll have been published as the *SRA Youth Inventory* by Science Research Associates, 57 West Grand Avenue, Chicago 10, Illinois. The *Manual* gives the technical data on reliability, item-test correlations, norms, etc.

Table 1

Percentage Differences between Low and High Socio-economic Status Groups on Problems that Predominate in the Lower Group

Note: all differences reliable at the 1% level of confidence or better.

Item No.	Socio-economic Status		Difference
	Low N=1809	High N=646	
19. I would like to have more vocational courses.	31	25	6
38. What shall I do after high school?	51	37	14
42. Should I go to college?	36	26	10
46. I can't afford college.	24	9	15
51. I must select a vocation that doesn't require college.	16	6	10
62. What jobs are open to high school graduates?	45	25	20
63. How do I go about finding a job?	37	30	7
95. I feel that I'm not as smart as other people.	36	27	9
101. I get stage fright when I speak before a group.	55	45	10
123. I wish I could carry on a pleasant conversation.	36	28	8
131. I want to learn to dance.	35	25	10
140. There aren't enough places for wholesome recreation where I live.	46	34	12
155. I can't find a part-time job to earn spending money.	30	20	10
157. I have no quiet place where I can study at home.	16	10	6
158. I can't get along with my brothers and sisters.	18	12	6
182. I wish I had my own room.	20	12	8
256. My teeth need attention.	19	7	12

relations between socio-economic status as measured and the incidence of problems checked by the low and the high socio-economic status groups. The items are the 39 that yielded such differences.

Beyond demonstrating validity of the status scale, the results summarized in Table 1 also give something of a qualitative picture of the kinds of preoccupations, worries and problems which characterize the low status group. By way of contrast it is of interest to examine the items that yield significantly higher proportions of responses from the high status group. They are shown in Table 2.

It is of interest to note that, if we take the averages of Low and High groups in Tables 1 and 2 as indices of the amount of worrying among teen-agers in the country as a whole, it is evident that the two groups worry about the same amount. They differ sharply, however, in the kinds of worries that they have.

### Summary and Conclusions

The Guttman test of scalability was applied to two independent random samples drawn from a total sample of approximately 10,000 high school pupils' responses. Validity of the scale was investigated in terms of significant differences between items in the *SRA Youth Inventory* and the socio-economic status index used in the *Purdue Opinion Panel*. The data support the following conclusions.

1. The items of the Socio-economic Status Index are scalable and represent substantially a unidimensional scale.

2. The scale is valid in that it correlates significantly with individual problems reported by a national sample of high school pupils.

3. The two groups into which the respondents are divided have about the same amount

Table 2

Percentage Differences between Low and High Socio-economic Status Groups on Problems that Predominate in the Upper Group

Note: all differences reliable at the 1% level of confidence or better.

Item No.	Socio-economic Status		Difference
	Low N=1809	High N=646	
9. I would like to take courses that are not offered in my school.	31	37	6
11. I have too much homework.	19	25	6
39. For what work am I best suited?	54	60	6
40. How much ability do I actually have?	58	64	6
41. I want to know more about what people do in college.	35	42	7
44. How shall I select a college?	34	49	15
126. I want to make new friends.	49	55	6
132. I have a desire to feel important to society or to my own group.	19	27	8
152. I'd like to know how to become a leader in my group.	21	27	6
154. I have difficulty budgeting my time.	19	28	9
169. I want to be accepted as a responsible person by my parents.	18	24	6
254. I don't get enough sleep.	14	20	6
277. How can I help get rid of intolerance?	12	23	11
278. How can I help to make the world a better place in which to live?	27	39	12
279. What can I do about the injustice all around us?	13	23	10
281. I'm worried about the next war.	29	37	8
282. Is there something I can do about race prejudice?	21	39	18
283. Is there any way of eliminating slums?	21	32	11
284. What can I do to help get better government?	13	20	7
285. How can I learn to use my leisure time wisely?	23	30	7
288. What can I contribute to civilization?	10	17	7
298. I wonder about the after life.	20	29	9

of worries, but these are qualitatively very different for the two groups.

Received December 19, 1952.

References

1. Gage, N. L. Scaling and factorial design in opinion poll analysis. *Studies in Higher Education LXI; Further Studies in Attitudes. Series X.* Division of Educational Reference, Purdue University, December, 1948.

2. Gage, N. L. and Remmers, H. H. Opinion polling with mark-sensed punch cards. *J. appl. Psychol.*, 1948, 32, 88-91.

3. Guttman, L. The Cornell technique for scale and intensity analysis. *Educ. psychol. Measmt.*, 1947, 7, 247-279.

4. Kerr, W. A. and Remmers, H. H. *The American Home Scale.* (Originally published by Science Research Associates, Chicago, Illinois, and to be republished by Psychometric Affiliates, Chicago Institute of Technology, Chicago, Illinois.)

5. Remmers, H. H. Measuring the public opinion of tomorrow. *Indiana Teacher*, May 1941, p. 281.

6. Remmers, H. H. The Purdue Opinion Poll for young people. *Scientific mon.*, 1945, 60, 292-300.

## The Relation of Motivation and Skill to Active and Passive Participation in the Group<sup>1</sup>

Ben Willerman

*Student Counseling Bureau, Office of the Dean of Students, University of Minnesota*

In any group there is variation in the extent of participation by different members in the activities of the group with some members more active than others. Among other factors, it is likely that differences in the personal characteristics of the members will help to explain the variation in extent of participation.

Many voluntary organizations have both a social and an organizational aspect with two corresponding areas in which group members may participate. Participation in the more purely social activities of the group may be quite different from working to recruit members, planning and arranging meetings, etc. This study is concerned with the latter type of participation.

The purpose of this study is to explore two general hypotheses concerning the differences between active members (AM) and passive members (PM):

1. AM as compared with PM are more motivated to participate in organizational activities, become more involved in the organization and derive more satisfaction from the organization. These differences are due, in part, to the possession to a greater degree by the AM group of personality characteristics which dispose them to become interested in and participate in groups.

2. AM more than PM have abilities which are probably related to effective action in organizations. In this study, it is assumed that the skills required to perform organizational functions are largely verbal and are ade-

quately measured by an academic aptitude test.

### Method

A questionnaire was given to 19 of the 20 academic sororities on the campus of the University of Minnesota. These questionnaires were completed by approximately 90% of the entire sorority membership. Two questions were used to select the subjects. The question used to select the AM of the sorority was, "List the names of the members of your sorority who would be a real loss to the sorority if they became inactive." The question used to select the PM was, "List the names of the members of your sorority who do not seem to have much interest in the sorority." To be included in the sample as an AM, a girl had to be selected by at least one-third of the responding members of her sorority. The corresponding criterion for PM was 10% or more.

In order to control two variables which were correlates of active and passive participation but not relevant to this study, some members of the sample were eliminated: (1) Girls living in the house were excluded because they were more frequent in the AM group and not frequent enough in the PM group to warrant separate analysis. The sample, therefore, consists entirely of town girls who generally commute from their homes to the University. (2) Girls who were members less than one year were excluded to eliminate the effects of a short membership period. This procedure left 41 AM and 37 PM. Almost all of the AM held an important sorority office. Few of the PM held an office. This outcome suggests that our dimension of active and passive membership is similar to but not identical with the leadership-followership dimension. The followership classification does not exclude active members who are not leaders nor does the leadership classification necessarily include active members who contribute to the

<sup>1</sup> This report is one of a series of research studies in student life being conducted by the Office of the Dean of Students, University of Minnesota. Various staff members gave helpful advice and assistance with this study. The author is grateful to Jack Laugon, St. Olaf College, formerly with the Student Activities Bureau, Office of the Dean of Students, University of Minnesota, for his assistance in a pilot phase of this investigation.

This study was supported in large part by a research grant from the Graduate School of the University of Minnesota.

group but not as leaders. Bird (1) and Stogdill (5) have excellent summaries of differences between leaders and non-leaders.

The questionnaire contained items related to the member's satisfactions and dissatisfactions with her sorority. Five-point rating scales elicited self-estimates of importance to the group, participation in the group, feelings of group belongingness, satisfaction with and acceptance of group decisions. Friendship choices and extra-sorority activities, as well as some background and demographical data were also obtained.

For many of the girls, test scores were on file at the Student Counseling Bureau. The tests included the Minnesota Multiphasic Personality Inventory (MMPI), the ACE Psychological Examination, and high school rank (HSR).

Chi<sup>2</sup> is the statistical test of significance used most frequently here and when used the distribution is split at the median.

Results

The AM derived more satisfaction, in general, from their membership than the PM. The first source of evidence for this result as well as for the "validation" of the selection method comes from five self-rating scales. Table 1 shows that the AM more than the PM believe the group regards them as im-

Table 1

Self-Ratings of Participation, Perceived Importance to Group, Group Belongingness, Agreement with Actions of Group, and Acceptance of Group Decisions when in Disagreement\*

	Active Members (N=41) %	Passive Members (N=37) %
Participate More or Much More in Sorority Activities than Most Members	93	5
Important and Very Important Real Part of Sorority	71	14
Agree Most of Time with Actions of Group	98	62
Complete Acceptance of Group Decisions	73	30
	49	11

\* All of these differences are significant well beyond the 5% level by Chi<sup>2</sup>.

portant, and consider themselves as participating in the sororities' activities more than most members. A larger proportion of AM express feelings of group belongingness, are satisfied with, and agree most of the time with the group's decisions.

The second type of evidence relates more to the sources of satisfaction and dissatisfaction than to the amount of satisfaction. The free answers to the questions, "What do you like most . . . ?" and "What do you least like about your sorority?" were coded using categories defined largely in terms of the actual answers. Coding agreement between two coders based upon 64 questionnaires for the "like most" answers was almost 75% and for the "like least" answers, based upon 34 cases, was about 80%. Since coding in some individual categories was very unreliable, to conserve space no table of these results is presented and only outstanding differences will be discussed.

Two main differences occur. The AM girls like the "group spirit, cooperation, and unity" of their sorority more than the PM (37% vs. 5%, Chi<sup>2</sup>, P < .05). The AM also more frequently like least the "lack of interest or cooperation of some members" (51% vs. 5%, P < .05). The PM more frequently like least the "compulsory functions" (16% vs. 0%, Fisher's exact test for 2 x 2 tables, P < .05). Although the difference is significant at only the 10% level, the PM more frequently complain that the sorority takes too much of their time. We may infer from these differences that the AM derive their satisfactions and dissatisfactions from the attainments or frustrations of the organizational goals, while the PM are less oriented this way and seem to regard some features of the sorority as an interference with their personal life.

One specific hypothesis concerning the lower interest of the PM in the sorority organization was suggested by the staff members of the University's Student Counseling Bureau on the basis of their experience with the MMPI. Girls who had an intense interest in male companionship tended to have on the MMPI a relatively high psychopathic deviate (Pd) score and a relatively low masculinity-femininity (Mf) score (low on this scale means

more feminine). It was reasoned that this type of attraction to persons outside of the sorority would be associated with lowered interest in the sorority. On this hypothesis the discrepancy scores between the Pd and Mf scales were compared for the two groups with the expectation that the discrepancy scores for the PM would be greater than those for the AM. The results together with some other test scores are shown in Table 2. While the difference between the two groups is not significant at the conventional 5% level, the *t* test (two-tailed) is at approximately the 9% level. There is, then, the possibility that one of the causes of low participation in the sorority group is the strong interest in men. Another interpretation is that when irresponsibility or non-conformity (high Pd) is coupled with a low interest in organizational functions (with which femininity may be correlated), the result is a girl neither adept at nor interested in the tasks of maintaining an organization.

If active participation in a group is the result of more than just situational circumstances, we should expect AM to be more active in other organizations. While we do not have a direct measure of amount of time or effort devoted to other organizations we do have evidence concerning the number of memberships held. This measure which is probably related to active participations shows that the AM belong to more extra-sorority groups ( $\chi^2$ ,  $P < .01$ ).

The greater participation of the AM in other organizations does not seem to be limited to the formal aspects of participation. In response to the question asking who their best friends at the University were, the AM mentioned 2.00 and the PM mentioned 1.09 girls on the average who were not members of the sorority ( $\chi^2$ ,  $P < .05$ ).

Logically, passive participation may result from either low motivation in the direction of participation or, if motivation is present, from counter-tendencies which oppose participation. The latter may be labelled "restraints" against participating. A plausible type of restraint in social situations is "fear of failure" or lack of self-confidence. To test the possibility that the PM possess characteristics which block participation, the K scores of the MMPI were compared for the two groups. The justification for the use of K as a measure of self-confidence rests upon an inspection of the items, many of which have "face validity," and upon the opinion of some counselors who believe that K often reflects genuine self-confidence rather than defensiveness. The results are consistent with the above reasoning. The mean K score for the PM is significantly lower than for the AM.

Another condition which may prevent an individual from contributing to a group is lack of skills demanded by organizational tasks. In a sorority the organizational tasks seem to require a relatively high level of abstract ability and skill in communicating. A

Table 2  
Scores on HSR, ACE, and MMPI\* for Active and Passive Members

Measure	Active Members			Passive Members			<i>t</i>	P
	N	Mean†	SD	N	Mean†	SD		
HSR	37	83.9	16.2	36	64.6	25.9	3.97‡	<.01
ACE	37	64.9	25.6	35	44.7	25.6	3.34	<.01
K	31	60.3	7.9	26	53.4	7.7	3.34	<.01
Pd	31	52.8	8.5	26	55.8	7.9	1.36	>.10
Mf	31	48.8	7.0	26	46.2	9.9	1.14	>.10
Pd-Mf	31	4.1	11.0	26	9.6	13.0	1.75	>.10 >.05
Sie	18§	43.4	7.2	16§	55.0	10.2	3.85	<.01

\* Data of other scales of MMPI not reported because they were not regarded as relevant to this study.

† Mean percentile scores for HSR and ACE; mean T scores for MMPI scale.

‡ Critical ratio was used because of unequal variances.

§ The N's for this scale are fewer than for other scales because some subjects' tests were scored before the Sie scale was constructed.

girl low in verbal ability might either be discouraged from participating by her fellow members or impose restraints upon herself because of fear of failure. Another possibility is that the necessity for maintaining a minimum grade average would leave a girl of low academic ability little time for organizational activities.

Although we do not have sufficient evidence to choose among these alternative hypotheses, a comparison of the mean scores of the ACE (total) indicates that skill as a determining factor is an effective variable. Table 2 shows that a difference of 20 points in ACE exists between the two groups. In addition, the HSR, a measure of past academic achievement, gives similar results.

Since many of the girls in the PM group had high ACE scores and many had low Pd-Mf discrepancy scores, the question was raised as to why these girls were passive participants. The answer to this question may be found in the correlation between these two sets of scores which for the PM is  $r = .45$  ( $N = 25$ ,  $P < .05$ ). Thus, the more a PM has an indicator related to active participation (high ACE), the more she tends to have an indicator which is related to passive participation (high Pd-Mf discrepancy).

For the AM group the variance of ACE scores is lower and moreover no such relationship was predicted. The corresponding  $r$  of  $-.22$  is not significant.

The large and statistically significant differences between the two groups on the Social Introversion-Extroversion scale of the MMPI is perhaps further validation for the scale (2, 3, 4) and reinforces the hypothesis of the existence of personal factors producing differences in participation.

### Summary and Conclusions

The purpose of this study was to test the hypotheses that active and passive participation in the organizational functions of a group was related to motivation to belong to the particular group, to general tendencies to be oriented toward participating in groups, and to skill in performing the tasks required by the organization.

The nomination technique was used to select 41 girls who were active members and

37 girls who were passive members from a total of 19 college sororities. Self-rating scales, sociometrics and test scores provided the basic data.

1. The active members were more attracted to their groups and seemed to derive satisfactions and dissatisfactions more from the organizational features of the sorority than the passive members.

2. The active members belonged to more student organizations and had more friends outside the sorority, indicating a general tendency to be attracted to organizations and to be socially inclined more than the passive members.

3. Using the discrepancy score of the Pd minus Mf scales of the MMPI as an indicator of strong interest in men which detracts from affiliation with the sorority, the passive members turned out to have higher scores. However, an alternative hypothesis that this discrepancy score represents a combination of non-conformity (Pd) and absence of interest in organizations (Mf) is also plausible.

4. The lower K scores of the MMPI for the passive members are interpreted as the presence of lack of confidence which operates as a "restraint" against participating.

5. As measured by the Sie scale of the MMPI, the passive members are definitely more introverted, indicating a general tendency for personal factors determining participation in a particular group.

6. The active members were 20 points higher than the passive members on both the ACE and HSR. As indirect measures of aptitude and ability for organizational tasks these differences lend support to the hypothesis that skill is an important factor in participation.

Received December 3, 1952.

### References

1. Bird, C. *Social psychology*. New York: Century, 1940.
2. Drake, L. E. A Social I. E. scale for the MMPI. *J. appl. Psychol.*, 1946, 30, 51-54.
3. Drake, L. E. and Thiede, W. B. Further validation of the Social I. E. scale for the MMPI. *J. educ. Res.*, 1948, 41, 551-556.
4. Gough, H. A research note on the MMPI Social I. E. scale. *J. educ. Res.*, 1949, 43, 138-141.
5. Stogdill, R. M. Personal factors associated with leadership: A survey of the literature. *J. Psychol.*, 1948, 25, 35-71.

# Studies in Social Interaction: III. Effect of Variation in One Partner's Prestige on the Interaction of Observer Pairs \*

Bernard Mausner

*University of Massachusetts*

It is a basic assumption of contemporary advertising and politics that people's judgments can be swayed by the opinions of individuals with high prestige. Clear-cut proof or disproof of this assumption in field situations is difficult. For example, an attempt to test the effects of endorsements by sports heroes on the sale of cigarettes would probably be beset by many problems. It may be of interest therefore to test a parallel assumption in the laboratory. The effect of one individual's judgments on those of another has been studied extensively through the use of an experimental design first employed by Sherif (9). In this, judgments are made first by each *S* alone, then in a group situation. Some of this work has indicated that the degree to which any one *S* will abandon his own judgment range for that of a partner will depend, in part, on the characteristics of that partner (3, 4, 5).

The present study was designed to test the effect of variation in a partner's prestige on social interaction in such an experimental design. Art judgments were chosen because these can be demonstrated to be stable in the absence of social stimulation, and because the manipulation of prestige in the area of art is relatively simple. The following experimental hypothesis was tested: *Ss* will be influenced more by the judgments of a partner with high prestige than by those of one with low prestige.

## Method

A group of forty undergraduate students in Washington Square College was given the Allport-Vernon Scale of Values. From these were chosen three groups of ten *Ss* matched for their percentile scores on the scale of aesthetic value (1). All of these *Ss* were

\* Formerly at Washington Square College, New York University. The writer wishes to acknowledge the assistance of Mr. Frank Celentano in the conduct of the experiment. Mr. Ted Climis played the roles of "fellow student" and "art director."

then given the Meier Art Judgment Test individually. Except for the omission of information concerning differences between the paired plates, standard instructions were followed. In a second session one week later the *Ss* were told that they were to repeat the test in order to determine its reliability. *Ss* in Group I (control) repeated the test substantially as in the first situation. *Ss* in Groups II and III were told that, to save time, the test would be given to pairs. A confederate of the experimenter was the second member of each pair. He was introduced to *Ss* in Group II as a fellow-student, to *Ss* in Group III as the art director of a local advertising agency interested in the results of the test.

In the together situation both members of the pair judged each of the pairs of pictures; the *S* was first in all even trials, the confederate in all odd trials. The confederate had memorized the test; he consistently stated the preference indicated as "wrong" by the scoring key.

## Results

The degree of social influence is expressed in terms of the change in the number of wrong preferences from the first to the second session. Table 1 gives mean wrong judgments

Table 1

Mean Wrong Choices on the Meier Art Judgment Test  
*M<sub>a</sub>* Gives Results for Pretest *M<sub>b</sub>*: Alone for Group I, with "Fellow Student" for Group II, and with "Art Authority" for Group III

Note: *S* first on even trials, partner first on odd.

	Group I		Group II		Group III	
	Odd	Even	Odd	Even	Odd	Even
<i>M<sub>a</sub></i>	14.6	13.6	13.3	13.7	10.9	16.1
<i>M<sub>b</sub></i>	14.8	13.9	17.0	14.5	19.6	18.2
<i>t<sub>a,b</sub></i>	.10	.11	3.9	.53	13.0	1.3
<i>p</i>	>.8	>.8	<.01	>.6	<.001	>.2
<i>N</i>	10	10	10	10	10	10

for both sessions for each group with the odd and even trials treated separately. Also included are  $t$ 's for the differences between mean wrong judgments alone and with a partner.<sup>1</sup>

Group I showed no significant change in the frequency of wrong judgments from situation one to situation two. This would be anticipated from the high reliabilities reported for this test: coefficients of reliability range from .70 to .85 (7). Groups II and III showed no significant change in the frequency of wrong judgments for the even trials where S made his choice first. It can be assumed therefore that there was no tendency for association with a partner of such deplorable taste to affect the general ability of S to discriminate good from bad pictures. However, in the odd trials, where the partner made his choice first, both Groups II and III demonstrated an increase in the frequency of wrong judgments (cf. Table 1). The differences are significant at the .01 level for Group II, at far better than the .001 level for Group III.

An examination of the data for individual Ss reveals that in Group II two of the ten Ss showed fewer wrong responses in the second situation than in the first. All of the Ss in Group III gave more wrong responses in the social than in the individual situations. Interpretation of differences between Groups II and III is justified to the extent that they are drawn from a relatively homogeneous population, and were matched for aesthetic values. The mean increase in wrong responses is significantly greater in Group III than in Group II:  $M_{DII} = -3.8$ ,  $M_{DIII} = -7.2$ ,  $t = 1.70$ ,  $n = 9$ ,  $p$  approaches .05 when the difference is evaluated in terms of a one-tailed hypothesis. This hypothesis is considered valid because the experimental hypothesis presented did not involve the probability of decrease in wrong responses, and because it was predicted that Group III would show more shift than Group II. Certainly the difference between a  $t$  of 3.9 for Group II and 13.0 for Group III would

indicate that the changes should be more reliable for the latter even though there is no way of comparing the two  $t$  values quantitatively. These results demonstrate that the judgments of a partner affected the responses of Ss taking the Meier Art Judgment Test, and that Ss tended to converge more consistently towards a partner with high prestige (art director, Group III) than towards one with little prestige (fellow student, Group II).

### Discussion

The writer has suggested (6) that the group judgment situation may be considered to create conflict for S between a tendency to continue giving his prior judgments, and one to agree with his partner. The findings of the present study indicate that the latter tendency may be affected by the partner's prestige. This does not deny the role of other factors in determining the extent of group influence. The effect of expert opinion on other kinds of judgment has been extensively investigated with, unfortunately, no consistent results; in some work "experts" were less effective in shifting judgment than "ordinary people," in others more effective (8, pp. 946-980). The differences among these reports may be due on the one hand to unreported or unanalyzed differences in other determinants of interaction: the stimulus factor, the nature of the response, the personalities of the Ss themselves. On the other hand, this variation in the reported results may be due to an inadequate specification of the nature of prestige (2). The teacher may be prestigious in some areas, not in others. The "expert" may exert maximum influence only when his "expertness" is accepted by S as genuine. Prestige, thus, might have to be measured in terms of influence on judgment.

One possible way of avoiding this circularity would be to attempt to relate various independent determinants of prestige to degree of convergence between members of a coacting pair of Ss. These could include status in a hierarchy, group membership, or past history of contact between Ss. In the present study prestige is varied by means of instructions to S regarding his partner's group membership (expert vs. non-expert). The

<sup>1</sup> For detailed table showing frequency of wrong judgments by each S in each situation order Document 3917 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting \$1.25 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.25 for photoprint readable without optical aid.

results indicate at least that under relatively controlled conditions it is possible to produce consistent variation in the degree of social interaction as a function of prestige manipulated in this manner. In further investigations attempts will be made to vary the prestige factor more extensively through variation in other determinants.

From a practical point of view, the present findings, with an obvious extrapolation, support the reliance on "expert testimonial" which has long been accepted practice in business and politics. However, the above discussion indicates that a caution is necessary. It is not always possible as yet to predict when an externally labelled authority will actually be accepted as an authority and will be able to exercise appreciable influence on judgment.

#### Summary and Conclusions

In a test of the effect of variation in one partner's prestige on the interaction of observer pairs three groups of ten Ss, equated for interest in art by means of the Allport-Vernon Scale of Values, were given the Meier Art Judgment Test. Ss in Group I repeated the test alone; Ss in Group II and Group III repeated it with a partner. He was introduced to Group II as a fellow student, to Group III Ss as an "art authority." The partner in all cases made choices indicated as wrong by the scoring key.

Degree of social influence was measured in terms of the shift in frequency of wrong judgments from the "alone" to the "social" situation. Group I (control) showed no significant shift in mean number of wrong judgments.

Both Groups II and III showed an increase significant for Group II at the .01 level, for Group III beyond the .001 level. Comparison of the two groups using a one-tailed test of significance shows Group III (art authority) giving a greater increase of wrong judgments than Group II (fellow student). This difference approaches significance at the .05 level.

It is concluded that the judgments of Ss taking the Meier Art Judgment Test were affected by the responses of coacting partners, and that this effect was a positive function of the partner's prestige.

Received November 21, 1952.

#### References

1. Allport, G. W., Vernon, P. E., and Lindzey, G. *A study of values*. Cambridge: Houghton Mifflin, 1951.
2. Asch, S. E. The doctrine of suggestion, prestige, and imitation in social psychology. *Psychol. Rev.*, 1948, 55, 250-277.
3. Berenda, Ruth W. *The influence of the group on the judgments of children*. New York: King's Crown Press, 1950.
4. Bovard, E. W., Jr. Group structure and perception. *J. abnorm. soc. Psychol.*, 1951, 46, 398-405.
5. Bray, D. W. The prediction of behavior from two attitude scales. *J. abnorm. soc. Psychol.*, 1950, 45, 64-84.
6. Mausner, B. Studies in social interaction: I. A conceptual scheme. *J. soc. Psychol.*, in press.
7. Meier, N. C. *The Meier art tests. I. Art judgment. Examiner's manual*. Iowa City: State University of Iowa, 1942.
8. Murphy, G., Murphy, L. B., and Newcomb, T. *Experimental social psychology*. (Rev. Ed.) New York: Harpers, 1937.
9. Sherif, M. A study of some social factors in perception. *Arch. Psychol.*, 1935, No. 187.

## Methods of Conducting Critiques of Group Problem-Solving Performance

E. Paul Torrance

*Human Resources Research Laboratories Detachment, Stead Air Force Base, Nevada*

The purpose of this study is to evaluate the relative effectiveness of four alternative methods for conducting brief critiques of a short problem-solving exercise designed to assist groups (air crews) to function more effectively as groups.

In many training situations, both military and civilian, it is necessary to conduct brief on-the-spot critiques of a group's performance. Instructors of the Advanced Strategic Air Command Survival School, the scene of the present study, are faced with this problem many times during the course of the field training of each crew they instruct. In all of these situations, there is the problem of how much guidance by the instructor or expert produces the best results. Can a crew effectively criticize itself and improve its problem-solving performance, or is the assistance of the expert necessary? When the expert conducts the critique, should he be the evaluator or should he keep the locus of evaluation within the crew?

### Theoretical Considerations

Much has been written in the areas of counseling and guidance and industrial training about techniques applied to the individual to bring about proper evaluation and improved adjustment or performance. One set of considerations deals with the locus of evaluation. One group, of which Rogers is the chief spokesman, holds that only when the locus of evaluation is in the individual does real growth and development take place (20). According to this theory, an evaluation by an expert or an evaluation resulting from a test would remove the locus of evaluation from the individual and would not result in development and growth. Essentially the same theory is represented in the work of Cantor (1, 2), Maier (14), Lippitt (12), French (4), Katzell (10), Haas (5), and others.

If one were to apply this theory to the

problem of critiques, the superior method would be expected to be one in which the leader assumes a non-evaluative role and stimulates the group to evaluate its performance and discover improved methods.

A second set of considerations centers about the role of group decision in changing behavior. Recent findings in industrial research and nutritive education research (6, 8, 11) indicate that group discussion as such results in very little change in behavior, while group decision as a component of group discussion brings about considerable change. In these experiments, scientifically developed information was given by the expert as it was needed but the decision was left to the group. Haire (6) points out, however, that group decision does not work with passive or apathetic groups, although its use almost always stimulates a desire for participation and eventually changes the apathy.

A number of experiments have explored situations and leadership techniques which set up resistance or retard growth, and others which win acceptance or stimulate growth. The problems of resistance have been treated by Zander (22), Torrance (20) and Coch and French (3). All emphasize the importance of respecting the individuals or groups involved. A variety of methods are discussed by Maier (14, 15, 16), Cantor (1, 2), Haas (5), Haire (6), Lippitt (12), and Rogers (18). There seems to be agreement that improved performance does not result merely through reading or hearing lectures. More active participation methods, such as through discussion and role playing procedures, are required.

The skill of the leader must also be considered as a factor. A series of experiments conducted by Maier (17, p. 170) showed that "a leader, if skilled and possessing ideas, can conduct a discussion so as to obtain a quality of problem-solving that surpasses that of a

group working with a less skilled leader and without creative ideas. Further, he can obtain a higher degree of acceptance than a less skilled person."

Maier concludes, however, that "even an unskilled leader can achieve good quality solutions and a high degree of acceptance" using democratic leadership. In another experiment (16), he demonstrated the superiority of the permissive discussion leader over the self-critique discussion with an observer present. Maier maintains that the major part of the difference was due to the relatively greater influence exerted by individuals with minority opinions in the "leader" groups than in the "observer" groups. "A discussion leader can function to up-grade the group's thinking by permitting an individual with a minority opinion time for discussion" (16, p. 287).

### Method and Procedure

*Subjects.* The subjects of the experiment were 57 combat air crews undergoing training at the Strategic Air Command's Advanced Survival School at Stead Air Force Base, Nevada. Most of these crews were B-29 (11 men) crews, but a few B-50 (10 men) and B-36 (usually about 15 men) crews were also included. Most of the crews had been functioning as crews for about four months, although some had been together for two or more years.

*Problem-Solving Exercises.* Two of the Intellectual Talents Tests (401-B and 701-X) developed by the Human Resources Research Laboratories were used. Both tests are thought to tap common-sense judgment and are alike in that each presents the examinee with problem-situations too complex for solution by any step-by-step logical reasoning process and requires the examinee to select the most essential or most critical of the many elements presented in the problem-situation. The problem-situations are rather commonplace and can be solved on the basis of knowledge gained from background experiences common in most persons' lives. Differences in the 401-B and the 701-X are that the 701-X consists of a larger number of shorter problems and permits an unlimited number of choices.

*Experimental Procedures.* The crews were tested in tents measuring 16 feet by 32 feet on the first day of their training. Each crew was first given an orientation regarding the nature and purpose of the test. Following this, each member of the crew was asked to make an estimate of his crew's performance. The first problem-solving test was then administered, after

which a post-test estimate of crew performance was obtained from each crew member.

Following this, a critique of the first problem-solving performance was conducted by one of the following methods:

1. Unstructured non-authoritarian or crew-centered critique: The crew was asked to evaluate and discuss its own performance. Discussion was centered on both the decision as to method and the way it was reached, as well as the way the decision was executed. The experimenter tried to stimulate discussion and encourage crew members to evaluate their performance, but the experimenter did not evaluate their performance. The experimenter accepted questions but referred them back to the crew. The attitude of the experimenter was definitely non-authoritarian. Techniques used were similar to those described by Cantor (1, 2), Maier (14), and Rogers (19).

2. Directive or expert critique: The experimenter diagnosed the performance of the crew according to a set of 13 rating scales (listed later), pointed out ineffective procedures, and suggested ways of improvement. He stated that through research, certain characteristics have been found to differentiate between crews which operate effectively and those which do not. The analysis included both the way the group went about making its decision and what they decided, as well as how they worked together to carry out the decision.

The experimenter took a very active role, assuming the role of the "expert." He tried, however, to give his advice in the most tactful way possible. He, nonetheless, gave definite evaluations and advice. The experimenter accepted questions and answered them as an "expert."

3. No critique: The experimenter went ahead and administered the California F-Scale which required about 15 minutes, before administering the second problem-solving test.

4. Self-critique: Time was allotted for a critique and the experimenter left the tent, returning after 15 minutes.

5. Structured non-authoritarian or crew-centered critique: The experimenter used the set of rating scales as a guide in getting the crew to evaluate itself and discover more effective ways of performing. The locus of evaluation was still within the crew, however.

Following the 15 minute critique period, the second problem-solving test, the 701-X, was administered. The rules were the same as for the first problem except that the time limit was ten minutes.

*Observations and Ratings.* After each of the two problem-solving tests, the experimenters completed a set of five-point rating scales following a set of descriptive scales on each of the following characteristics: (1) Organization of manpower; (2) Selective use of personnel; (3) Supervision; (4) Participation in decision-making;

(5) Acceptance of suggestions or criticisms; (6) Consideration of available time; (7) Checking work; (8) Leadership function; (9) Survey of the situation; (10) Understanding instructions; (11) Group atmosphere; (12) Speed of reaction to the problem situation; and (13) Officer-airmen relations.

### Results

A problem-solving score was computed for each crew on both of the problem-solving tests, using the scoring formulae already in use for these tests. A performance rating was also computed for each crew on both of the problem-solving situations by adding the thirteen ratings made by the examiner. In order to hold constant scores and ratings for the first problem-solving test and to determine if the variance in scores and ratings is due to the method of conducting the critique, analyses of co-variance were then carried out both for ratings and for scores. Using the ratings, the variance for critique methods was found to be statistically significant at the one per cent level of confidence ( $F = 4.968$ ). Using problem-solving scores, however, the variance was not statistically significant at less than the five per cent level of confidence ( $F = 1.957$ ). Because of the small number of crews critiqued by each experimenter by each method, it was not possible to compute the interaction of experimenter and critique method.

Crews participating in the unstructured non-authoritarian critique were combined with those participating in the self-critique and crews participating in the expert critique were combined with those participating in the structured non-authoritarian critique in order to study the effect of structure vs. non-structure

in critiques. Analysis of co-variance revealed that the variance due to structure is significant at the five per cent level both for ratings ( $F = 5.664$ ) and for scores ( $F = 5.124$ ). Analysis of co-variance also showed that the variance due to different experimenters is not statistically significant ( $F = 0.429$ ) for ratings and for scores.

In order to study relative improvement in performance which might be attributable to differences in methods of conducting critiques, each crew was ranked in order from one to fifty-seven on each of the four variables (score on 401-B, score on 701-X, ratings on 401-B performance, ratings on 701-X performance). Crews were then divided equally into a most improvement category and a least improvement category on ratings and on scores. Table 1 shows the percentage falling into each category according to method of conducting the critique for both ratings and scores.

The *t*-test of significance of differences in percentage reveals the superiority of the expert critique over the non-authoritarian critique (significant at the .001 level of confidence), no critique (significant at the .01 level), and the self-critique (significant at the .02 level). The differences in percentages between the expert critique and the structured non-authoritarian critique is not statistically significant. The latter tends to be more frequently followed by improvement than are the unstructured non-authoritarian critique (significant at the .01 level of confidence), no critique (significant at about the .10 level of confidence), and the self-critique (not statistically significant).

The situation in regard to improvement on problem-solving scores is about the same as

Table 1  
Comparison of Effectiveness of Methods of Conducting Critiques

Basis of Comparison	Expert Critique (11 crews)	Structured Non-Authoritarian Critique (11 crews)	Self-Critique (12 crews)	Unstructured Non-Authoritarian Critique (11 crews)	No Critique (12 crews)
Percentage showing "most improvement" in standing on scores	73	73	33	36	33
Percentage showing "most improvement" in standing on ratings	91	64	50	9	33

for ratings, except that the superiority of the expert critique is not as clear. The *t*-test of significance of the difference in percentage shows that the expert and structured non-authoritarian methods are superior to the unstructured non-authoritarian, the self-critique and no critique at about the .02 level of confidence. The unstructured non-authoritarian method and the self-critique appear to have no superiority over no critique.

### Discussion

The fact that the structured non-authoritarian is superior to the unstructured non-authoritarian method and that the expert method is not superior to the structured non-authoritarian method would suggest that the locus of evaluation is not important in the type of critique studied in this experiment. Of course, it may be that even though the "expert" makes evaluations, the crew still makes its own evaluations and does not surrender its evaluative function to the expert as readily as some might suppose. A close examination of crews subjected to the expert method and making little improvement indicates that some of the evaluations given by the "expert" were definitely rejected by the crew. The crucial thing may be the giving of evaluations that can be accepted rather than the giving or not giving of evaluations.

The issue of group decisions does not become crucial in this experiment since in every case the decision was left to the crew, although that decision may have been made by one person, usually the aircraft commander. In using the unstructured non-authoritarian, however, it was observed by almost all of the experimenters that a crew would recognize and discuss improved solutions and even appear to give general approval to these solutions. Yet, when the time came to decide how to organize for the second problem, the Aircraft Commander would simply say, "We'll do it the same way we did the other one." This may explain why this method is not more effective than no critique of any kind.

In regard to the overcoming of resistance, the less structured methods are least effective. It must be mentioned, however, that some of the crews which made the most outstanding

improvement were crews using the self-critique. The difficulty is that not all crews are able to look objectively at their performance and discover more effective ways of working together. Most crews seem to require enough structure or guidance to assure that their evaluations and considerations will be concerned with the salient elements. This does not in any way deny the importance of the participation and involvement of the group. It does, however, emphasize the importance of the "expert" and the nature of the role he must play in order to be effective where single trial, immediate performance is concerned.

Although the variation due to experimenter differences was not significant, differences in the success of experimenters were observed. For example, 70 per cent of the crews critiqued by two of the experimenters were in the "most improvement" category while only 25 per cent of the crews of another experimenter were in this category (significantly different at about the 5 per cent level of confidence). The least well trained experimenter differed very little from the best trained experimenters.

The results would appear to have important implications for training of many types, especially training of the on-the-job variety in industry, education, and the military services. Although there are a number of questions which need to be subjected to further study, the results of this study seem to point the way to using structured critiques where decisions are still left to the group, where final evaluation is left to the group, but where the trainer can help guide the evaluative process. This study also suggests several directions for further research which are being pursued through a series of additional studies now under way. These studies are concerned with the role of the expert, the decision-making techniques of the group's usual leader, spread of learning within the group, and transfer of learning to more different situations.

### Summary

A total of 57 combat air crews undergoing survival training were divided randomly into four experimental groups and one control group. Each experimental group was administered a problem-solving test, critiqued ac-

cording to one of four methods, and then administered a second problem-solving test. The control group was given no critique between the two problem-solving tests.

Crews obtained scores on both of the problem-solving tests and ratings of manner of performance on both of the tests.

Analysis of covariance indicates statistically significant variances in ratings due to method of conducting critiques. Analysis of covariance indicates statistically significant variance in both scores and ratings due to structuring the critique but no statistically significant variance due to experimenters.

Crews critiqued according to the more highly structured methods are more frequently followed by "greater improvement" than are crews critiqued according to the less highly structured methods. Crews participating in the unstructured non-authoritarian and the self-critique do not perform significantly better than crews receiving no critique.

Received November 24, 1952.

### References

1. Cantor, N. *The dynamics of learning*. Buffalo, N. Y.: Foster and Stewart, 1946.
2. Cantor, N. *Learning through discussion*. Buffalo, N. Y.: Human Relations for Industry, 1951.
3. Coch, L. and French, J. R. P. Overcoming resistance to change. *Human Relat.*, 1948, 1, 512-532.
4. French, J. R. P. Field experiments: changing group productivity. In J. G. Miller (Ed.), *Experiments in social process*. New York: McGraw-Hill, 1950.
5. Haas, R. B. Action counseling and process analysis; a psychodramatic approach. *Psychodrama Monogr.*, 1948, No. 25.
6. Haire, M. Some problems of industrial training. *J. soc. Issues*, 1948, 4, 41-47.
7. Hendry, C. E., Lippitt, R., and Zander, A. Reality practice as educational method. *Psychodrama Monogr.*, 1947, No. 9.
8. Hendry, C. E. *A decade of group work*. New York: Association Press, 1948.
9. Human Resources Research Laboratories. *Intellectual Talents Tests 701-X and 401-B*. Washington, D. C.: HRRL, Bolling Air Force Base.
10. Katzell, R. A. Testing a training program in human relations. *Personnel Psychol.*, 1948, 1, 319-329.
11. Lewin, K. Group decision and social change. In T. M. Newcomb and E. L. Hartley (Eds.), *Readings in social psychology*. New York: Henry Holt and Co., 1947.
12. Lippitt, R. An experimental study of the effect of democratic and authoritarian atmospheres. *Univ. of Iowa Studies in Child Welfare*, 1940, 16, 43-195.
13. Lippitt, R. *Training in community relations*. New York: Harper, 1949.
14. Maier, N. R. F. *Principles of human relations*. New York: John Wiley, 1952.
15. Maier, N. R. F. and Zerkoff, L. F. MRP: a technique for training large groups of supervisors and its potential use in social research. *Human Relat.*, 1952, 5, 177-186.
16. Maier, N. R. F. and Solem, A. R. The contribution of a discussion leader to the quality of group thinking: the effective use of minority opinions. *Human Relat.*, 1952, 5, 277-288.
17. Maier, N. R. F. The quality of group decisions as influenced by the discussion leader. *Human Relat.*, 1950, 3, 155-174.
18. Rogers, C. R. *Client-centered therapy*. Boston: Houghton Mifflin Co., 1951.
19. Rogers, C. R. Divergent trends in methods of improving adjustment. *Harvard Educ. Rev.*, 1948, 38, 209-219.
20. Torrance, E. P. The phenomenon of resistance in learning. *J. abnorm. soc. Psychol.*, 1950, 45, 592-597.
21. Torrance, E. P. and Levi, M. *Crew performance in a test situation as a predictor of performance in the field*. Stead Air Force Base, Nevada: HRRL Detachment No. 3, 1952, in press.
22. Zander, A. Resistance to change—its analysis and prevention. *Advanc. Mgmt.*, 1950, 15, 9-11.

## Logical Reasoning: With and Without Training \*

William J. Morgan and Antonia Bell Morgan

*Aptitude Associates, Merrifield, Virginia*

It is very strange indeed that psychologists have paid so little attention to problems of logical reasoning. During the last 50 years no systematic and comprehensive approach has been undertaken by them toward these problems. We made a careful search of the literature since 1927 and found 21 references to experimental studies of logical reasoning, and we were rather generous in our interpretation of what constitutes an experimental study. Generally speaking, therefore, psychologists have been contributing less than one experimental study per year on problems of logical reasoning. But they continue to speculate and philosophize, not quite so often as the philosophers themselves, on the characteristics of this mental process. It is difficult to understand why supposedly hard-bitten, scientifically-minded psychologists have given so little attention to this problem. Perhaps in their failure to exploit the findings of Störing (7, 8, 9) and Eidens (2), they became discouraged. Psychologists seem to be under the delusion that logical reasoning is confined to the syllogism, a view which has long been abandoned by the logicians themselves.

It may also be that psychologists have not been willing to undertake experimental studies of logical reasoning, because they were, for such a long time, desperately trying to divorce themselves from the influence of philosophy, and, of course, logic is an integral discipline of philosophy. Whatever may be the reasons for the paucity of experimental studies of logical reasoning, it seems to be a fact that mathematicians rather than psychologists are concerned themselves with logic. In spite of the stress laid on logical reasoning, especially the deductive aspects, by Professor Clark Hull in his establishment of a behavior system, psychologists seem content to believe that logic, like any other game or sport, is free to set up its own rules of how the game

will be played. But unlike chess, or poker, or basketball, the rules of logic are basic to science. As H. M. Johnson, himself a psychologist, has said (3, p. 74) "No artful manipulation of symbols according to prescribed rules can make good logic out of bad logic . . . . The *structure* of science as we know it is predetermined by the definitions, postulates, and rules of manipulation of symbols that we call modern logic. This logic includes the whole of the traditional or Aristotelian logic, cleared of certain well known defects; it includes a great deal that Aristotle and his imitators overlooked. . . . we may be sure that if any procedure assumes equivocation, affirming the consequent, denying the antecedent to be valid, then it does not yield a set of rules for 'scientific inference.' "

In view of the absence of too much experimentation on logic by psychologists, it is not surprising to find cropping up some rather far-fetched notions about the nature of logical reasoning. In the chapter on *Speech and Language* in Stevens' *Handbook of Experimental Psychology* (4, p. 806), Professor G. A. Miller says: "The fact is that logic is a formal system, just as arithmetic is a formal system, and to expect untrained subjects to think logically is much the same as to expect preschool children to know the multiplication table."

This sentence, an argument by analogy, contains a number of interesting implications which might be subjected to analysis but we shall restrict ourselves to the assertion contained therein that *untrained subjects cannot be expected to think logically*.

When we use the word "logic" we accept the definition given by Warren's *Dictionary of Psychology* (10) where logic is defined as the "principles that enable an individual to make judgments or conclusions which are consistent with the data at hand."

### Subjects and Procedures

The Morgan Test of Logical Reasoning (5) was administered to 134 adults, all employed

\* This paper was presented before Section I, Psychology, of the American Association for the Advancement of Science at its annual meeting, in St. Louis, Missouri, 30 December 1952.

by the United States Government. This test, which was first developed in 1946 for the testing of superior adults, contains 75 true-false items in verbal form. The scoring formula is Right minus Wrong. The subjects in this study were allowed 30 minutes. These are a few sample items from the test:

(a) All highly successful businessmen are practical psychologists. *Therefore*, some practical psychologists are highly successful businessmen.

(b) Most executives are college graduates. The majority of executives are Republicans. *Therefore*, most college graduates are Republicans.

(c) If we rearm Germany, the French will oppose us, and if we fail to maintain air bases in East Anglia we shall incur the resentment of the British. But it is essential to retain the good will of either France or Britain. *Therefore*, we must maintain our East Anglian air bases or else abandon plans for the rearmament of Germany.

(d) No person interested in treating human ailments has failed to study Professor Pavlov's book on the nature of the digestive juices—a book that won the Nobel prize. No person who has failed to study Professor Pavlov's book is a physician. *Therefore*, although they may have other interests, it can be said that all physicians are interested in treating human ailments.

(e) Many women are high-strung and emotional. A high-strung and emotional temperament is frequently a barrier to clear and logical reasoning. *Therefore*, many women are unable to reason logically.

(f) You can fool some of the people all the time. You can fool all the people some of the time. *Therefore*, you cannot fool all the people all the time.

The subjects consisted of two groups (WL and WOL) of 67 each. All subjects in Group WL (With Logic) had had at least three semester hours of college training in logic. No subjects in Group WOL (Without Logic) had had any training in logic. Each person in Group WL was paired in terms of sex, age, and college degree(s) with a person in Group WOL.<sup>1</sup> In each group there were 58 males and 9 females with a mean age of 27 years

<sup>1</sup> The pairs were matched in terms of educational achievement as measured by college degrees, rather than in terms of scholastic ability. However, for 61 cases in the group with logical training and for 65 cases in the group without logical training we have statistics derived from the Verbal Intelligence Test (published by Aptitude Associates, Merrifield, Va., copyright 1948). This test is scored by summing the rights, and the maximum score is 50. The mean score for Group WL was 38.3 (SD, 7.7); and the mean for Group WOL was 32.1 (SD, 9.5). The Critical Ratio ( $D/\sigma_D$ ) was 4.0. This difference is significant at the one per cent level.

and a standard deviation of 5.0. The oldest was 42, youngest 20, median 27. There were in each group 43 with a Bachelor's, 16 with a Master's, and 8 with an LL.B. degree (plus the Bachelor's). In addition to Groups WL and WOL, there were 9 subjects with a Ph.D. degree, 7 males and 2 females. The oldest was 57, the youngest 26, the median 33, and the mean age 36 years. None of the Ph.D.'s had had any training in logic.

### Results

The lowest score in Group WL was -2, the highest 67, mean 29.1, and the standard deviation 14.0. The lowest score in Group WOL was -7, the highest 48, mean 21.2, and the standard deviation 11.2. The means were significantly different beyond the 1% level.

By comparing the mean score for Group WL with the mean score for Group WOL, it is found that Group WOL did 73 per cent as well as Group WL. Since this test is scored Right minus Wrong, if a person is guessing throughout, he should get a zero score, all other things being equal. If a person does not know how to reason logically, he would have to guess on this test on every item, and we would expect him to get a zero score. But what do we find? Instead of a zero score, these college graduates who did not have the benefit of formal training in logic were actually able to achieve a mean score of 21.2 compared with a mean score of 29.1 for those who had had training in logic. In other words they did 73% as well as the group which had received training in logic. This is a far cry from zero.

Although Group WL obtained a higher mean score on the test than Group WOL, it is remarkable that of the LL.B.'s, 7 of the 8 who had not received training in logic obtained higher scores than their paired partners; of the Master's, 6 of the 16 who had not received training in logic did better on the test than their paired partners; of the Bachelor's, 13 of the 43 who had not received training in logic obtained higher scores than their opposite numbers in the WL Group. In other words, 26 of the 67 subjects, i.e., 38%, in the WOL Group did better than their paired partners who had received training in logic.

The lowest score for the Ph.D.'s was 23, the highest score 45, mean 32.7, and the standard deviation 7.4. There was far less variability

in scores in the Ph.D. Group by comparison with either Group WL or Group WOL. The mean score for the Ph.D.'s is significantly higher at the 1% level than the mean score for Group WOL. The mean score for the Ph.D.'s is also higher than the mean score for Group WL, and the chances are 88 out of 100 that the difference is significant.

### Conclusions

1. In the majority of cases, college graduates who have had at least three semester hours of college training in logic obtain higher scores on a test of logical reasoning than college graduates who have not had courses in logic.

2. Professor Miller's hypothesis that untrained subjects cannot be expected to think logically is not substantiated, however, because: (a) 38% of the subjects who had had no training in logic obtained higher scores than their paired partners who had had college training in logic; (b) subjects without college training in logic did 73% as well, as a group, on the test of logical reasoning as those who had had training in logic; and (c) college graduates with Ph.D. degrees who had not had college courses in logic obtained higher scores, as a group, than college graduates with B.A., M.A., and LL.B. degrees who had had courses in logic.

3. Professor Miller's hypothesis might be restated to read, "In the majority of cases, untrained subjects cannot be expected to think as logically as trained subjects."

### Further Implications

There are two problems that need to be explored by further research: (1) Are students with facility in clear thinking the ones who are usually attracted to courses in logic? and (2) To what extent is proficiency in logical reasoning the result of formal courses in logic rather than attributable to other factors, such as the subject's native intelligence? (11).

We have found in other studies that scores on tests of logical reasoning always correlated positively, often substantially, and sometimes very highly with scores on group tests of intelligence such as the Henmon-Nelson, the Thurstone ACE, the Miller Analogies, and the Verbal Intelligence Test. Other investigators such as Wilkins (12, p. 28), Burt (1, p. 237), and Sells (6, p. 23) have obtained similar re-

sults. We are, therefore, inclined to suggest that the ability to think logically is, to a certain degree, an aspect of intelligence.<sup>2</sup>

We believe that the potentiality for learning to reason logically is dependent upon the native intelligence of the individual, and the rules by which logical reasoning is governed are learned in the daily experiences of life, sometimes in the classroom, with or without the benefit of instruction in formal logic. It would be desirable to find out what experiences and courses, other than formal courses in logic, increase the student's proficiency in logical reasoning. It is our opinion that some courses, such as mathematics, even though not labeled as courses in logic, may have considerable "carry-over" value to logical reasoning.

Received December 31, 1952.

### References

1. Burt, C. L. *Mental and scholastic tests*. London: P. S. King & Son Co., 1921.
2. Eidens, H. Experimentelle untersuchungen über den denkverlauf bei unmittelbaren folgerungen. *Arch. f. d. ges. Psychol.*, 1929, 71, 1-66.
3. Johnson, H. M. If-then relations in paralogics. *Psychol. Rev.*, 1944, 51, 69-75.
4. Miller, G. A. Speech and language. Chapter 21, pp. 789-810, in Stevens, S. S., *Handbook of experimental psychology*. New York: John Wiley & Sons, 1951.
5. Morgan, W. J. and Morgan, A. B. *The Morgan Test of Logical Reasoning*. To be published about June 1954 by Aptitude Associates, Merrifield, Virginia.
6. Sells, S. B. The atmosphere effect. *Arch. of Psychol.*, 1936, 200, 1-72.
7. Störing, G. Experimentelle untersuchungen über einfache schlussprozesse. *Arch. f. d. ges. Psychol.*, 1908, 11, 1-127.
8. Störing, G. Psychologie der disjunktiven und hypothetischen urteile und schlusse. *Arch. f. d. ges. Psychol.*, 1925, 54, 23-84.
9. Störing, G. Psychologie der zweiten und dritten schlussfigur und allgemeine gesetzmässigkeiten der schlussprozessen. *Arch. f. d. ges. Psychol.*, 1926, 55, 47-110.
10. Warren, H. C. *Dictionary of psychology*. Boston: Houghton Mifflin Co., 1934.
11. White, E. E. A study of the possibility of improving habits in thought in school children by a training in logic. *Brit. J. educ. Psychol.*, 1936, 6, 267-273.
12. Wilkins, M. C. The effect of changed material on ability to do formal syllogistic reasoning. *Arch. of Psychol.*, 1928, 102, 1-83.

<sup>2</sup> It would be desirable to cross-validate this study, with special care devoted to matching WL and WOL Groups of college graduates on the basis of measured intelligence.

## The Effect on Recall of Changing the Position of a Radio Advertisement \*

William A. Belson

*Birkenbeck College, London, England*

This inquiry was aimed at establishing the relative rates of recall, under normal listening conditions, of a given advertisement placed (a) at the beginning of a program (beginning advertisement) and (b) in the middle of a program (interruption advertisement). It was, in effect, an inquiry into the relative rates of recall of a beginning advertisement (B) and an interruption advertisement (I). This comparison was made in respect of normal (N) or at-home listening conditions and not in respect of the situation in which people listen with the intention of learning. Whatever the relative merits of the two advertisements in the latter situation, there are plausible grounds for theorizing that under N listening conditions various subjective evocations such as hostility, inattention and certain defense mechanisms tend to enter effectively into the perception processes to the special detriment of I.

The grounds for this theory are two-fold. In the first place, the interruption type of advertisement emerged, in a preceding survey in Sydney, as the most disliked form of radio advertising and, indeed, as a source of no little hostility (2). Second, the work of Bartlett, Levine and Murphy, Rapaport and others (1, 7, 9, 10) has indicated the importance to perception of affective tendencies, partisanship and various personal factors.

While, however, a superiority of B over I would conform to the theory, such a comparison would not provide a crucial test. Differences in recall between B and I could, in fact, arise out of conditions other than differential

subjective evocations. To provide a crucial test, it was necessary to examine differences in recall of B and I first where those subjective evocations peculiar to the N situation were operative and secondly where they were eliminated. The latter situation required, in fact, the development of a learning set (L) in relation to B and I.

### Method

*Design.* Four groups G1, G2, G3 and G4 were matched according to intelligence, age, sex, occupation and general background. Two of them, G1 and G2, heard B and I (respectively) under N conditions. This meant that there was an attempt to evoke in them N reactions (i.e., NB and NI). The other two matched groups, G3 and G4, heard B and I (respectively) after the development in them of L. Each of the four groups was subsequently tested for recall of the advertisement to which it had been exposed. Difference in recall (R) between G1 and G2 (i.e., "NB - "NI) represents the advantage in recall of one placement over the other. The full extent of the difference in recall which may be attributed to differential N reactions is equal to ("NB - "NI) - ("LB - "LI).

It is conceivable, however, that such differences in recall as might occur could arise out of unplanned group differences in respect of personal characteristics and testing conditions. To provide a check on this possibility a control device was incorporated into the design. Advertisements B and I were carried by identical programs, though of course they were recorded on different wires. Sections of additional advertisement were introduced in equivalent positions on *each* wire. These two additional sections constituted the control material on each wire and each group was also tested for recall of this control material. If unplanned differences had not occurred, then differences in recall of control interest in either the N or the L situations should not be significant. Details of this control device are presented in Figure 1.

*Material.* Material used included two wire recordings of commercial programs, playback equipment, program opinion sheets and question booklets.

Two Wire Recordings of Commercial Programs. The advertisements were carried by the

\* While the inquiry is presented here in comparative isolation, it was in fact conducted as part of a wider investigation into the relation of attitude to recall in radio advertising. Findings from this wider study will be introduced only where they contribute to the interpretation of the present results. The investigation was carried out in Sydney, Australia, where commercial broadcasting predominates (2). The author is now studying for the Ph.D. at Birkenbeck College.

Beginning Placement of the Advertisement		Interruption Placement of the Advertisement	
Part Control Material	We have pleasure in presenting to you the story of "Sherry and Son," brought to you by Raymonds, the makers of distinctive <i>sweets</i> . Have you tasted Crunch Block, the sweet with the <i>special flavour</i> ? It's made by Raymonds, the sweet makers of distinction. Only the <i>best ingredients</i> go into it— <i>honey, glucose and nuts</i> —all of them <i>ideal for sweets</i> . Crunch Block is <i>really worth tasting</i> and is <i>obtainable at the manufacturer's own store in the Royal Arcade and at confectioners, grocers and milk-bars</i> .	We have pleasure in presenting to you the story of "Sherry and Son," brought to you by Raymonds, the makers of distinctive <i>sweets</i> . Have you tasted Crunch Block, the sweet with the <i>special flavour</i> ? It's made by Raymonds, the sweet makers of distinction. Only the <i>best ingredients</i> go into it— <i>honey, glucose and nuts</i> —all of them <i>ideal for sweets</i> . Crunch Block is <i>really worth tasting</i> and is <i>obtainable at the manufacturer's own store in the Royal Arcade and at confectioners, grocers and milk-bars</i> .	Part Control Material
	Raymonds, the makers of Crunch Block, have developed a <i>new process</i> called <i>Bubbling</i> which makes the texture of the sweet <i>fine and smooth</i> . This product is <i>vitamin packed</i> and has <i>special nutritive value</i> and is manufactured under strictly <i>hygienic conditions</i> . Crunch Block <i>costs threepence</i> and there is <i>no shortage of supply</i> .	FIRST HALF OF "SHERRY AND SON"	
B	"SHERRY AND SON" (ALL)	And now we briefly interrupt our story. Raymonds, the makers of Crunch Block, have developed a <i>new process</i> called <i>Bubbling</i> which makes the texture of the sweet <i>fine and smooth</i> . This product is <i>vitamin packed</i> and has <i>special nutritive value</i> and is manufactured under strictly <i>hygienic conditions</i> . Crunch Block <i>costs threepence</i> and there is <i>no shortage of supply</i> .	I
	This program is brought to you by Raymonds, the makers of Crunch Block, the <i>sweet of distinction</i> . Don't forget to ask for it; you'll enjoy its special flavour.	SECOND HALF OF "SHERRY AND SON"	
Part Control Material		This program is brought to you by Raymonds, the makers of Crunch Block, the <i>sweet of distinction</i> . Don't forget to ask for it; you'll enjoy its special flavour.	Part Control Material

FIG. 1. Text of the two advertisements and their positions relative to program and control material. Items on which recall was tested are italicized.

program "Sherry and Son." As shown in Figure 1, the advertisements fell into several parts. On each wire there was an opening advertisement. On one wire, however, half the opening advertisement had been shifted to the middle of "Sherry and Son." This was the I placement. On the other wire the equivalent section of the advertisement was not moved and it was this placement which was called B. Hence it will be seen that as far as placement was concerned, the two advertisements had a preceding and a following statement in common, and it was this additional material which constituted the control. On the other hand, the experimental material was B/I.

Program Opinion Sheet. This was a single sheet which asked for written opinions of each

of three programs. It was part of the technique used in deceiving subjects into reacting normally to the advertisement. Details follow.

**Instructions.** N. Situation. The wire carrying B was played to G1 and that carrying I to G2. Subjects were told that the purpose of the session was to get their opinions of these programs and were provided with opinion sheets for this purpose. These programs were said to be taken direct from the library of one of the commercial radio stations and to be just as they would be if they were going on the air. The first of the programs, a collection of three vocal items called "Just for You," was played with its advertisement; the playback machine was stopped and subjects were asked to write their candid opinions of the program. The purpose of this

Table 1  
Recall Scores\* on Beginning and Interruption Advertisements

Conditions of Exposure	Beginning Advertisement		Interruption Advertisement		Significance of Difference†	
	Mean	SD	Mean	SD	CR	P
Normal Reaction	3.42	2.86	1.88	2.17	2.46	0.016
Learning Set	4.91	2.38	6.67	2.87	2.26	0.032

\* Score out of 16 marks.

† Two tails of the distribution.

was to facilitate the deception. When subjects had finished this, they heard "Sherry and Son" with its advertisement and subsequently wrote opinions of that program too. After this the question booklets were distributed; there had been no warning at all of this step. Subjects were asked to write down their feelings (like/dislike) about the advertisement in "Sherry and Son" and about advertising in general. They were then given recall tests by specific question<sup>1</sup> and multiple choice methods.

**L. Situation.** The wire carrying B was played to G3 and that carrying I to G4. Subjects were told that while they would be asked for opinions of the programs, their main job was to listen to the advertisement in "Sherry and Son" and that they would be required to recall it at the end of the program. They were asked to concentrate on remembering the advertisements and to keep out of the picture any attitude they may have towards radio advertising in general or towards this particular advertisement. Opinions of the programs were asked for and recall tests were made as with G1 and G2.

**Scoring.** Items on which scores were based are those italicized in Figure 1. Marks were given for each correct reproduction, one on the specific question system and one on the multiple choice system, making a total of 16 marks on

<sup>1</sup> What was the name of the product being advertised? What were the contents—the ingredients—of the product? I mean what did it have in it?

the 8 items included in the B/I material and a possible of 28 marks on the 14 items included in the control material. Marks for the recall of the names of producer and product were excluded from totals because these items were common to the B/I and the control material.

### Results

From Tables 1, 2, and 3 it will be seen that in the N situation recall of I is very significantly *less than* recall of B ( $P = .016$ ), whereas in the L situation, recall of I is very significantly *greater than* recall of B ( $P = .032$ ). This represents a large and significant reversal of the advantage of I in going from the L to the N situation ( $P = .002$ ). Expressed in terms of percentage of recall, B was recalled in the N situation about twice as well as I (21% *vs.* 12%), whereas in the L situation B was recalled only three quarters as well as I (31% *vs.* 42%).

This phenomenon does not, for the following reasons, appear to be an artifact arising out of unplanned differences between groups in respect of personal characteristics or testing conditions. First, recall of control material occurring with B and I, respectively,

Table 2  
Recall Scores\* on Control Material Occurring with the Beginning and Interruption Advertisements, Respectively

Conditions of Exposure	Beginning Advertisement†		Interruption Advertisement†		Significance of Difference	
	Mean	SD	Mean	SD	CR	P
Normal Reaction	7.92	2.90	8.26	4.24	0.40	0.62
Learning Set	10.75	4.76	11.42	3.40	0.55	0.57

\* Score out of 28 marks.

† Control material occurring with the experimental placement.

Table 3  
Characteristics of Matched Groups\*

Group	Intelligence†		Age		Sex	Size of Group
	Mean	SD	Mean	SD		
G1	5.04	0.81	26.40	2.26	male	31
G2	5.30	0.71	25.10	2.00	male	41
G3	5.09	0.87	25.17	2.44	male	24
G4	5.30	0.82	25.00	2.60	male	24

\* The occupation and background of the four groups were the same: trainee carpenters under the Commonwealth Reconstruction Training Scheme; ex-servicemen (non-commissioned) recruited from semi- or unskilled occupations.

† In standard scores.

does not differ significantly in either the N ( $P = .64$ ) or the L ( $P = .57$ ) situations, while the slight advantage in the L situation of control material occurring with B is maintained in the N situation ( $P = .79$ ). Secondly, the four groups appeared to be well matched and test conditions did not noticeably deviate from plan.

Moreover, this reversal phenomenon was repeated with each of the 8 items in B/I on which recall tests were made.

#### Discussion

It is not difficult to theorize about the causes of the relative disadvantage of the I placement. A theory of inhibition through hostility would not only have a certain plausibility, but would also be supported by the fact, already reported, that in an accompanying survey in Sydney (2), the interruption placement emerged as the most disliked form of radio advertising. Some caution is needed, however, for it was also found (2) that verbalized attitude (in terms of like/dislike), whatever its concomitant organic processes might be, was not correlated with recall ( $r = +.01 \pm .11$  with intelligence partialled out). Under the circumstances, there is some case for suggesting the existence of a generalized defense mechanism of an involuntary type—a theory which gains additional support from a further finding (2) of very little or no correlation between alleged degree of attention to the advertisement and recall ( $r = -.14 \pm .11$  with intelligence partialled out). Quite clearly, however, further theorizing and research are required at this point.

A second interpretative point must be made. The present use of "captive audiences" leaves out of account certain aspects of the real home listening situation: people listening at home are usually free, during the broadcast of an advertisement, to walk about, talk or turn the set off. In fact, one of the claimed advantages of the interruption placement of an advertisement is that people are less likely to walk about, tune out, etc., during the middle of a program than at the beginning. The present study was not, however, directly concerned with this issue, although the issue is one on which research might well be conducted.

#### Summary and Conclusions

The prime purpose of this inquiry was to compare, under normal listening conditions, rates of recall of an advertisement placed at the beginning and in the middle of a program. There was, in fact, a good case for theorizing that reactions such as hostility, inattention and defense mechanisms would normally enter the perception processes to the special detriment of the interruption type of advertisement.

While a superiority in recall of the beginning advertisement would concur with this theory of differential reaction, a crucial test would also require a comparison of recall rates when such reactions were eliminated. Hence the full investigation involved a comparison of recall of the two advertisements after (a) normal reaction and (b) the establishment of a learning set.

Four matched groups were exposed in pairs under conditions (a) and (b), respectively, to a specially designed advertisement, one of each pair hearing the beginning placement and the others the interruption placement. A fairly elaborate administrative procedure was used to evoke "normal" reactions to the advertisements. A control device was designed to detect differences in recall arising out of unplanned group differences in respect of personal characteristics and testing conditions.

Results showed that normal reaction to the advertisement in the interruption placement interfered with perception much more than did normal reaction to it in the beginning placement. The difference was, in fact, such that the very significant advantage of the interruption advertisement under learning set conditions of exposure was reversed, and very significantly so, when normal reactions took place.

*Received November 24, 1952.*

## References

1. Bartlett, F. C. *Remembering: a study in experimental and social psychology*. Cambridge, England: Cambridge University Press, 1932.
2. Belson, W. A. *The relation of attitude to perception and recall in radio advertising*. Unpublished Bachelor's thesis, Univ. Sydney, 1949.
3. Chein, I. Behavior theory and behavior of attitudes: some critical comments. *Psychol. Rev.*, 1948, 55, 175-188.
4. Doob, L. W. The behavior of attitudes. *Psychol. Rev.*, 1947, 54, 135-156.
5. Droba, D. D. The nature of attitude. *J. soc. Psychol.*, 1933, 4, 444-463.
6. Garrett, H. E. *Statistics in psychology and education*. (3rd Ed.) New York: Longmans, Green and Co., 1947.
7. Levine, J. M. and Murphy, G. The learning and forgetting of controversial material. *J. abnorm. soc. Psychol.*, 1943, 38, 507-517.
8. McNemar, Q. Opinion-attitude methodology. *Psychol. Bull.*, 1946, 43, 289-374.
9. Rapaport, D. Emotions and memory. *Psychol. Rev.*, 1943, 50, 234-243.
10. Sellemann, V. The influence of attitude upon the remembering of pictorial material. *Arch. Psychol.*, N. Y., 1940, 258, 1-63.

## Check-Reading as a Function of Pointer Symmetry and Uniform Alignment<sup>1</sup>

Keith W. Johnsgard

*The State College of Washington*

During the past several years an extensive program of psychological research has been carried on dealing with human reactions to aircraft instrument panels. This program, undertaken primarily by the United States Air Force, is an attempt to simplify the increasingly complex task of reading aircraft instruments under flight conditions.

Aircraft instruments serve for three basic types of reading: (1) check-reading for assurance of a normal indication; (2) qualitative reading for the meaning of a deviation; and (3) quantitative reading for the actual numerical value of an indication (5). This paper is concerned with the first type and employs the rotating pointer type indicator which has been shown to facilitate short latency responses with a minimum of errors (3).

A recent study has indicated that rectangular arrangement of small engine instruments on multi-engine aircraft and the use of rotatable dials, making possible uniform pointer alignment under flight conditions, will provide a significant advantage in speed and accuracy of check-reading (6). However, another project concerning recognition span made it apparent that check-reading is facilitated by pointer symmetry even when the pointers are not all in the same standard position such as 9 o'clock (8). Both uniform pointer alignment and pointer symmetry are superior to mixed alignment for check-reading. The development of rotatable dials makes these principles applicable to engine instrument panels. With this arrangement dials could be fixed in such a manner that the pointers would form any pattern that would facilitate rapid and accurate check-reading wherein any deviation could be quickly identified in terms of direction, engine, and function.

<sup>1</sup> Submitted in partial fulfillment of the requirements for the M.S. degree in the Department of Psychology in the Graduate School of the University of North Dakota. The author wishes to thank Dr. Hermann F. Buegel under whose direction the research was conducted.

A basic problem then is a consideration of which type of pointer-pattern would facilitate the most efficient check-reading. It is this problem with which this paper is concerned.

### Method

*Stimulus Preparation.* Four sixteen-dial panels containing different pointer patterns were used and will be referred to as configurations throughout the remainder of this report. A null hypothesis was stated that the four configurations would equally facilitate check-reading. The configurations are shown with pointers in a null position in Figure 1.

Nineteen stimulus panels were prepared for each configuration. One panel showed the pointers in a null position with the remaining eighteen panels for a particular configuration containing dials in which pointers were deviating from null position. These eighteen panels were split into six blocks of three panels each. Panels of the first block each contained one deviating pointer, panels of the second containing two, and so on with each panel of the sixth block containing six deviating pointers. A complete set of eighteen test panels for any one configuration contained a total of 63 deviating pointers. The dial or dials

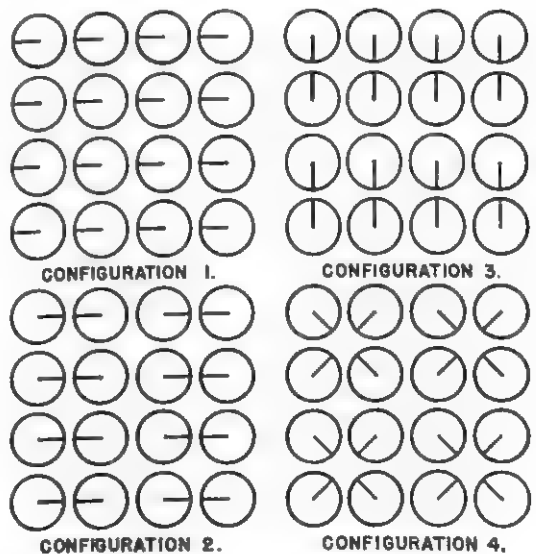


FIG. 1. The four configurations with pointers in a normal or correct position.

within a particular panel that were to contain deviating pointers were chosen in a random manner (4). The position of the deviating pointer within an error dial was chosen in the same way requiring a possible discrimination ranging from a maximum of 180 degrees to a minimum of 15 degrees from the null pointer position.

Stimulus material was prepared from 35 mm. negative film with projected dial borders and pointers appearing white on a dark background when flashed on a screen (2, 11). The diameter of a projected dial was three inches (7, 9, 10, 13). Pointer width was approximately  $3/32$  inch extending from the center of the dial to the border (1).

**Physical Conditions.** A modification of the Whipple pendulum type tachistoscope was used. The noiseless mechanism is described and pictured in a recent publication (11). Stimulus materials were contained in a slide projector behind the tachistoscope. The projector was 34 inches from the floor with the light beam projected horizontally to a screen seven feet away. The illuminated screen when using transparent slides was approximately eight foot-lamberts (2).

Two S's were tested simultaneously, and were seated on each side of the projected beam in two single-arm writing desks pointed directly at the center of the projected panel. The distance from the projection area to the S's eyes was approximately 50 inches (1).

The testing room was darkened to maintain effective contrast (14). Directly over the subjects a soft beam of light was directed downward to facilitate writing responses on score sheets. The beam did not affect screen brightness.

An exposure time of a half second was allowed. This is the average fixation time for pilots engaged in instrument flying (8).

**The Sample.** The sample population consisted of 48 male students enrolled at the University of North Dakota. The ages ranged from 18 through 37, with all but 7 of the S's between the ages of 18 through 26. Visual acuity was checked before experimentation with a Snellen Eye Chart. A criterion of 20/25 was set as a minimum of visual acuity. Subjects with sight corrected to this criterion by glasses were considered satisfactory for experimental purposes. None of the S's had experienced any tachistoscopic training of any kind.

**Test Procedure.** After being checked for eyesight, the two S's were seated, acquainted with response sheets, and instructed. The response sheets contained 72 sixteen-dial panels numbered in the same manner as the test panels. Before being presented with a set of 18 test panels the S's were allowed to study the projected normal configuration panel with all needles in a null position. This normal panel was then shown three times with a half second exposure. It was explained that on the test panels to follow not all of the dials contained pointers in the normal po-

sition. S's were instructed to check the appropriate dial or dials on the response sheets that corresponded to those on the test panel containing deviating pointers. Presentation of the 18 test panels for the appropriate configuration followed. Before an individual panel presentation the S was informed of the panel number and was given a ready signal. Approximately 1 second later the exposure occurred. After each exposure as much time as was needed was allowed for responding.

Following observation of two sets of test panels a short rest was allowed. The entire test period varied from 35 to 45 minutes depending on speed of response.

In an attempt to eliminate practice effect from the total group results the order in which the four tests were administered was varied. One group of 12 S's observed the configurations in numerical order. A second group began with configuration 2, a third group with configuration 3, and a fourth with configuration 4. S's for each group were selected at random.

**Method of Scoring.** One point was allowed for each dial correctly identified as containing a deviating pointer. Four total configuration scores were computed for each paper. Each of these total scores was the sum of six sub-scores. The sub-scores were the correct responses made to each of the six sets of three panels that contained from one to six deviating pointers. It was not considered necessary to penalize incorrect responses. The method was arbitrary.

## Results

Means of total correct responses in locating error dials in the entire set of 18 slides for each configuration are listed in Table 1. To test the null hypothesis that the configurations were of equal difficulty, a small sample *t* test for correlated means was computed. The results of this test are indicated in Table 2. The test showed configuration 3 to be significantly superior to the others tested for check-reading. The stated null hypothesis might safely be rejected. Differences between

Table 1  
Means, Standard Deviations, and Standard Errors  
for Total Correct Responses

Configuration	Mean	s	SE
1	30.19	6.80	.98
2	31.46	5.72	.83
3	34.48	4.85	.70
4	17.54	5.94	.86

Table 2

Confidence Levels and *t*-Scores Between Total Correct Response Means

Configuration Means Compared	<i>t</i>	Confidence Level
C1-C2	1.61	.120
C1-C3	4.94	Beyond .001
C1-C4	12.82	Beyond .001
C2-C3	3.92	Beyond .001
C2-C4	13.53	Beyond .001
C3-C4	18.35	Beyond .001

the standard deviations proved to be insignificant.

The performance curves for the 48 S's on each configuration are shown in Figure 2. The curves are best fit by a method of least squares and in all cases the fits were very reasonable. It should be stated that a definite restriction exists for the interpretation of these data. The data are plotted as mean correct responses as a function of blocks of three trials. However, an experimental maximum is imposed by design, since only one error dial

is contained in each panel of the first block, two in the second block, and so on with six in the last block. This factor could have influenced the first two blocks of trials but perhaps affected no further blocks. With this restriction in mind, the data were plotted, as curve shape was considered to be important with regard to further practice. Examination of the formula for configuration 2 is significant in that the suggested asymptote is 3.14, while those for configurations 1 and 3 are 2.26 and 2.64, respectively. There exists the possibility that with further practice, configuration 2 might prove to be more useful than any of those tested.

It will be recalled that the order of configuration presentation was varied in order to compensate for possible over-all practice effect. With regard to this an analysis of total configuration scores was made. The mean score for the 48 Ss on the first configuration presented was 27.42, with the mean score for the last configuration presented being 29.56. It is evident that some transfer exists between configurations.

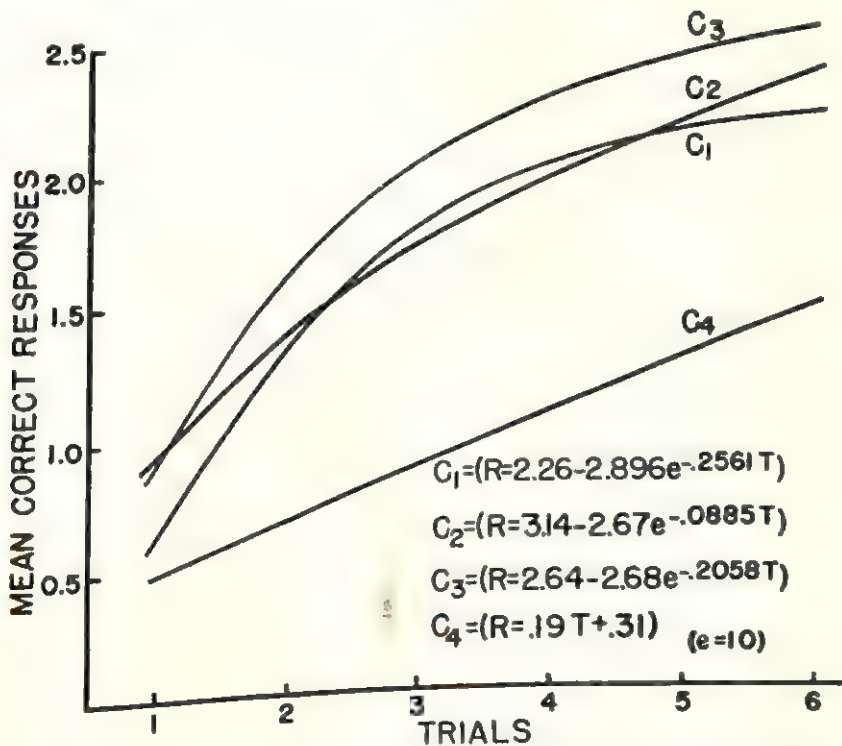


FIG. 2. Mean error dials located in each of the four configurations with each abscissa point representing the mean values of blocks of three trials.

Grether and Warrick (6) report that about twice as much time was required to check-read a sixteen-dial panel employing four sub-groups than to check-read a panel of equal dial number with pointers in uniform alignment at the nine o'clock position. This study tends to reinforce that finding in that about twice as many error dials were found in configuration 1 with uniform alignment at nine o'clock as in configuration 4 which employed the four sub-groups. These same investigators have shown that the nine o'clock position is the most favorable pointer position of uniform alignment. This experiment has indicated that panels employing pointer symmetry are equally as good as the most favorable position of uniform alignment for check-reading. The results further suggest with reasonable assurance, that one of the configurations (C3) with pointer symmetry is superior to uniform alignment (C1) and that another (C2) might prove superior with practice.

The findings of this experiment would indicate that rotatable dials and a sixteen-dial rectangular panel would facilitate check-reading of aircraft engine instruments. It is likely that these principles can be applied in most situations where a multi-engine arrangement exists such as industry where rapid accurate check-reading is a necessity.

### Summary

A tachistoscopic study in which simulated instrument dials were observed at short exposure was performed to determine efficiency in locating deviating dial pointers in four instrument panels employing the principles of uniform alignment, pointer symmetry, and sub-grouping of pointer pattern. A null hypothesis was stated that the four patterns would equally well facilitate check-reading. S's totalled 48 naive male students.

The results of the experimentation allow a statement of the following tentative conclusions:

1. Configurations in this experiment employing pointer symmetry facilitate check-reading equally as well as do panels with uniform alignment. There is reasonable evidence

that one of the former type is superior to the latter in terms of number of correct responses, and that another might prove superior with practice. The null hypothesis was rejected.

2. Panels employing pointer symmetry and uniform alignment are superior to sub-groups for check-reading.

3. Check-reading improves with a relatively short amount of practice and some transfer exists between panels with differing pointer positions.

4. It was suggested that the use of a rectangular sixteen-dial panel of aircraft engine instruments with rotatable dials would facilitate rapid check-reading, and that these principles might profitably be applied in other situations where multi-engine panels are used.

Received January 12, 1953.

### References

1. Armed Forces-NRC Vision Committee. *Standards to be employed in research on visual displays*. Ann Arbor: University of Michigan, 1947.
2. Chalmers, E. L., Goldstein, M., and Kappaul, W. E. *The effect of illumination on dial reading*. U. S. Air Force Air Materiel Command, A. F. Tech. Report No. 6021, 1950.
3. Connell, S. C. *Some variables affecting instrument check reading*. U. S. Air Force Air Materiel Command, A. F. Tech. Report No. 6024, August 1950.
4. Edwards, A. L. *Experimental design in psychological research*. Rhinehart and Company, Inc., New York, 1950, pp. 23-24.
5. Grether, W. F. *Discussion of pictorial versus symbolic aircraft instrument displays*. US AAF AMC, Engng. Div., Aero Med. Lab., TSEAA-694-8B, August 4, 1947. (Unclassified, English)
6. Grether, W. F. and Warrick, M. J. *The effect of pointer alignment on check-reading of engine instrument panels*. U. S. Army Air Force Headquarters, Air Materiel Command, Engineering Division, Memo Report No. MCREXD-694-17, 1948.
7. Grether, W. F. and Williams, Jr., A. C. *Speed and accuracy of dial reading as a function of dial diameter and angular spacing of scale divisions*. In P. M. Fitts (Ed.), *Psychological research on equipment design*. Washington, D. C.: U. S. Government Printing Office, 1947.
8. Johnson, T. G. *Recognition span and reading patterns in simulated instrument dial formations*. Unpublished Master's Thesis, Univer-

- sity of North Dakota, Grand Forks, N. Dak., 1950.
9. Kappauf, W. E. and Smith, W. M. *Design of instrument dials for maximum legibility: II. A preliminary experiment on dial size and graduation.* USAF Air Materiel Command Memorandum Report MCREXD-694-1N, 1948.
  10. Kappauf, W. E., Smith, W. M., and Bray, C. W. *Design of instrument dials for maximum legibility: II. Development of methodology and some preliminary results.* USAF Air Materiel Command Memorandum Report MCREXD-694-1L, 1947.
  11. Stevens, S. S. *Handbook of experimental psychology.* John Wiley and Sons, Inc., New York, 1951, p. 1301.
  12. Thurstone, L. L. *A factorial study of perception.* Chicago: U. of Chicago Press, 1944, pp. 36-37.
  13. White, W. J. *The effect of dial diameter on ocular movements and speed and accuracy of check-reading groups of simulated engine instruments.* USAF Air Materiel Command Technical Report 5826, 1949.
  14. Woodworth, R. S. *Experimental psychology.* New York: Henry Holt and Co., 1938, p. 688.

## Visual Performance as a Function of Low Photopic Brightness Levels \*

Milton L. Rock

*E. N. Hay & Associates, Inc., Philadelphia*

A systematic investigation of performance in visual tasks as a function of low photopic brightness levels is essential to expand our knowledge of adequate visual performance levels. Although there have been many studies under higher brightness levels (above 1 foot-lambert) the range between cone threshold and 1 foot-lambert has been relatively neglected. Stimulated by the needs of the last world war, interest in this region has been increasing and the time seems now appropriate for a systematic summary of information in this area. When new experiments are added to fill the gaps, this should give us a more nearly complete theoretical and practical knowledge of the problem. Senders' (27) summary and Rock's (25) annotated bibliography (sections B and C) concerning studies of visual acuity are comprehensive and should be consulted for a background in this general area. In this discussion visual acuity *per se* will not be considered a measure of performance and in most studies reported was a controlled variable.

A number of variables have been shown to be of importance in the investigation of visual performance. For example, defects of both the visual mechanism and the stimulus objects must be controlled. Ferree and Rand (11, 12) and Sheard (28) have indicated that ocular defects, presbyopia in their cases, require increased illumination for adequate performance. Tinker's (37) study on illegible print also indicates a need for higher bright-

ness levels required with this defective stimulus object. Tinker (36, 38) has shown that adaptation level of the eye has an enormous effect on performance and also on subjects' brightness level preferences. Investigations of the effect on visual performance of the quality of light have given conflicting results. Ferguson and McKellar (9), investigating binocular visual observations of a landolt ring at brightness levels below 1 foot-lambert, found that at 0.5 foot-lamberts the best performance was with red light, then amber, white, blue and the poorest, green. Many other studies such as that of Craik (7), who investigated legibility of different colored instrument markings at low illumination, have found green and blue to be inferior. Hartline (15) investigated the relative merits of lights of different wave-lengths in the airplane cockpit situation and found the measure of individual thresholds to be a good index of visual function at low intensity levels. McFarland (22) recommends that for cockpit use at night, no wave-length below 620 mm. (red-orange) be used. Some studies have reported on the effectiveness of performance under various wave-lengths of light. Spragg and Rock (30) investigating performance of reading airplane dials under four different wave-lengths and at two illumination levels (.01 and 0.1 foot-lamberts) found that within the range of colors and brightness studied dial reading performance showed no consistent relationship to the wave-length composition of illuminant. These results have been verified with performance of flying a link trainer under these same conditions.

\* This report is a condensation and partial revision of a doctoral dissertation, the original of which is on file in the library of the University of Rochester. The author is indebted to S. D. S. Spragg, University of Rochester, for his direction and guidance.

The experiments reported here were conducted as part of a program of research on human factors related to aircraft instrument lighting carried out on a research contract (W33-038 ac18317) between the University of Rochester and the Air Materiel Command, U. S. Air Forces. They have been reported in the following technical reports to the Aero Medical Laboratory of the Air Materiel Command: MCREXD-694-21 and TR 6040.

The types of performance criterion which investigators have used fall into three main classes: (a) speed of response, (b) accuracy of response, (c) physiological correlates. Cobb (5, 6) has investigated speed of vision as a function of illumination level. He employed various patterns associated with confusion patterns as stimulus objects and found

that a logarithmic relationship held for parallel bars between 1 and 100 foot-candles, but under more complicated conditions the relationship breaks down so that the expected gain in sensitivity due to increased intensity is not realized. Ferree and Rand (10) investigating speed of vision as a function of brightness with special reference to industrial situations found that on work of a factory type, involving important use of the eyes, speed of vision increased as brightness increased up to a maximum. Many studies of reading performance have used speed as a criterion; Tinker (32, 33, 35).

Accuracy, usually in conjunction with speed, has been used extensively in practical situations as a measure of performance. Typical studies are those of White, Britten, Ives, and Thompson's (41) study of ocular efficiency and fatigue among letter separators as a function of brightness level, and Weston and Taylor's (40) investigation of fine type-setting done by hand as a function of brightness level. Most of the threshold studies, however, are accuracy studies which do not involve speed of reaction, e.g., Hartline (15), Brown (3), Brown and Mize (4), Graham and Hunter (14), etc.

Luckiesh and Moss (17, 18, 19) employed such physiological correlates as blink rate, heart and pulse rate, metabolic ratios, etc., as performance criteria for readability. McFarland, Knehr and Berens (23) found that metabolic ratio and pulse rate are inadequate criteria for reading. Tinker's (34) numerous experiments throw doubt on the use of blink rate as a criterion.

From his consideration of the existing literature, the writer believes that in future experiments on visual performance: (a) visually screened subjects should be used so that they fall into a "normal" category; (b) the stimulus objects should be legible and above the resolution threshold of the eye at all brightnesses tested; (c) light quality should be controlled and specified; (d) light quantity should preferably be designated as brightness; and (e) performance criteria should be speed and/or accuracy with the possible use of certain physiological correlates in the case of fatigue studies.

In consideration of the above, and with the

aim of contributing experimental data to fill the gaps in our knowledge of visual performance as a function of low brightness levels, four representative visual tasks were chosen for investigation: (1) judgment of magnitude of an illusion; (2) motion threshold; (3) depth perception; and (4) a simple addition task. Each task was investigated under five brightness levels in the crucial range of .005 foot-lamberts (which is just slightly above the values usually stated as cone threshold, i.e., .002 to .004 F.-L.) to 1.00 foot-lambert.

#### Experiment I. Magnitude of the Müller-Lyer Effect

This first experiment is a study relating performance in the judgments of equality of the two lines of the Müller-Lyer figure to various low photopic levels of brightness. The Müller-Lyer, one of the best known visual geometric illusions, has many variations. The basic example is a figure consisting of two straight parallel lines of equal length. Each line is terminated at each end by two short oblique lines forming an angle whose apex is at the end of the major line. On one line the oblique lines extend back toward the center of the major line, on the other they extend away. The illusion, which is a potent one, consists in perceiving the latter line as longer than the former.

This illusion was selected as one of the visual tasks to be investigated here because it is representative of a general class of visual illusions and hence is a rather important visual perceptual task of judgment.

Many variables affect the judgment of visual illusions. Our past experience, associations, demands, desires, and more or less obscure influences may create illusions. The physical characteristics of the stimulus object are also of paramount importance. The location of the object in the visual field, the structure of the field, equivocal figures, the influence of angles, color, irradiation, and brightness contrast and lighting and shadows are just a few of the main variables causing illusions.

Our present interest is to investigate the relation of the magnitude of effect of an illusory figure to low photopic brightness levels. We want to answer the following two ques-

tions: Do geometric illusions increase in effect under low brightness levels as compared with ordinary illumination? Is there a critical value of illumination above which the magnitude of the illusion does not vary with brightness?

### Method

*Apparatus.* The general plan of the apparatus followed that employed by Spragg and Rock (29) in their studies of dial reading performance as related to low photopic illumination levels.

The subject was seated in a three-sided booth, approximately  $4 \times 4$  feet, facing the middle wall. The entire visual field was painted a matte black. Placed in the  $14 \times 11$  inch aperture of the front wall was the Müller-Lyer figure. The center of the figure was 28 inches from the subject's eyes and  $15^\circ$  below his horizontal line of regard. An adjustable head rest, mounted on a horizontal bar, served to keep the subject's head in a satisfactorily constant and comfortable position.

The stimulus object was a Müller-Lyer figure, with a 7.5 cm. stationary standard arrow-headed part. On the subject's right side was a sliding board on which there was a line with an arrow feather at one end. This side of the figure could be moved under the standard until its line was the desired length. The lines were 2 mm. in width, white on a black background. The obliques were 3 cm. and at a  $27^\circ$  angle. The variable stimulus was moved by the experimenter by means of a rack and pinion which could be varied by equal steps in a smooth manner.

The light sources were two 60 w. Mazda lamps in cans fitted with filters and aperture holders. An assembly consisted of a ground-glass square of heat-resistant glass, and a brass plate with a circular aperture drilled in the center. Voltage was maintained at a constant level by means of a variac, Model V-5MT, and a monitoring Weston A.C. Voltmeter, Model 433. The color temperature was in the neighborhood of  $2400^\circ$  K. Chosen levels of illumination were achieved by means of accurately drilled apertures in removable brass plates. All light sources had two ground-glass surfaces in the optical pathway to achieve high dispersion.

Data sheets were prepared in advance. These indicated the five levels of illumination with columns under each for recording the subject's responses, and a section for the subject's comments.

*Subjects.* Ten male subjects served as the experimental population. All were students at the University of Rochester (three graduates and seven undergraduates) and were all in their late teens or twenties in age. Subjects chosen were those who passed a rigorous visual screening,

using the Keystone Telebinocular. All subjects had: normal ophthalmoscopy, 20/20 visual acuity, monocularly and binocularly, at distance and near, without glasses, 80 per cent or better stereopsis, no vertical imbalance, less than 6 prism diopters physiological exophoria; less than 2 prism diopters of exophoria or esophoria at distance; and normal color vision.

*Procedure.* Each subject was allowed to become cone dark adapted (approximately ten minutes) before the illumination was turned on. With the stimulus object illuminated with .05 foot-lambert of brightness, the subject was given the instructions:

This is an experiment to determine the influence of varying brightnesses of illumination on the perception of the length of a line complicated so as to present an illusion. Yes, this is a very common illusion, but I want you to tell me when this line (pointing to the variable) *looks to you* to be equal to this stationary line. (With the variable much larger than the standard.) Now, the line is much larger than the standard. You are to say "Now" when the lines appear to you to be equal in length. (Decrease the size of the variable stimulus so that it is much smaller than the standard.) Now the line is smaller than the standard. Say "Now" when it appears to you that the lines are equal. Five practice trials were given.

On the formal trials each subject was first presented with the variable stimulus larger than the standard. The stimulus was decreased by the experimenter in successive steps of  $\frac{1}{2}$  mm. until the subject reported equality. Then the experimenter presented to the subject a stimulus smaller than the standard, and the stimulus was altered by successive increments until the subject reported he no longer perceived any difference. This is a modified method of limits called the method of equivalents, in which only points of equality are recorded, approached from the two possible directions. Ten responses at each illumination level for each subject were required, five ascending and five descending. It is realized that space errors are present in this method, but as we are interested in differences between performance at the various levels of brightness and since these errors are presumably constant, they should not bias the results.

The levels of illumination were chosen to encompass the critical range of low photopic brightness as found in the experiment of Spragg and Rock (29). The levels range from just above cone threshold to 1 foot-lambert. The values used were: .005, .01, .05, .1, and 1 foot-lamberts.

Brightness measurements were made with a Macbeth Illuminometer used in the subject's position and directed against a white square painted with the same paint as that of the stimulus object.

Since five levels of illumination were used it was necessary to employ balanced sequences of

brightness levels to control possible practice and fatigue effects. In changing from one level to another the subjects were given from five to ten minutes for adaptation.

Each subject was given a visual screening test on one day, and the entire series of judgments on another day.

### Results

The data of this experiment consist of judgments of equality of length of lines in the Müller-Lyer figure for ten subjects under five different brightness levels.

The mean errors of judgment are presented in Table 1 which shows for each subject the mean error, sigma, standard error, and per cent error of standard at each of the five brightness levels.

Inspection of Table 1 and Figure 1 shows that: (a) the effect of the illusion at the three higher brightness levels is considerably less than at the two lower levels; (b) above 0.05 foot-lamberts increasing brightness produces no significant improvement in performance; while (c) below 0.05 foot-lamberts decreasing brightness is clearly associated with poorer performance on this task.

Since our principal concern is with performance as a function of brightness, a *t*

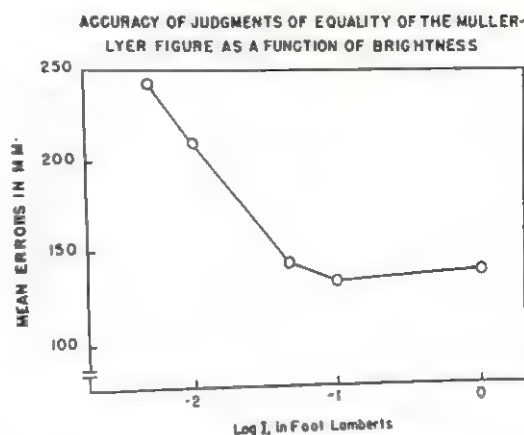


FIG. 1. Visual performance as a function of low photopic brightness levels.

analysis was carried out comparing group performance for each pair of brightness levels. A summary of this analysis is presented in Table 2. From Table 2 it is seen that all differences between brightness levels below .01 foot-lamberts and those above .01 foot-lamberts are significant at the 1 per cent level.

On the basis of the data presented above, it seems clear that: (1) there is a larger error in perception of length as tested in the Müller-Lyer figure below the region of .05 foot-lamberts; and (2) little or no increase in performance results from increasing brightness above this level.

**Practice Effects.** It will be recalled that each subject was given five practice trials before formal trials were begun, in order to reduce practice effects. The adaptation periods between levels would tend also to reduce practice effects. In order to determine whether practice effects were playing a significant role in this situation the errors were tabulated for each subject in terms of first brightness level tested, second level tested, etc. Since each brightness level appeared in each ordinal position the same number of times, no advantage due to sequence is present for any brightness level.

The *t* tests of the several differences indicate that there is no evidence of a practice effect. Early experimenters with illusions noticed that continued experience with one certain figure diminished the amount of the illusion. Heymans (42) and particularly Judd (42) made a systematic study of the

Table 1

Showing the Magnitude of Errors in mm. in Judgments of Equality of the Müller-Lyer Figure at Five Brightness Levels

Subjects	Brightness in Foot-Lamberts				
	0.005	0.01	0.05	0.10	1.00
1	2.4	1.7	.9	.8	.8
2	2.9	2.4	1.7	1.8	1.9
3	1.9	1.5	.9	1.0	1.0
4	3.7	3.8	2.8	2.8	2.6
5	1.9	1.5	1.0	1.0	1.1
6	2.7	2.7	2.2	2.2	2.3
7	3.0	2.4	1.0	.8	1.0
8	2.2	2.2	1.5	1.6	1.3
9	1.5	1.3	.9	.4	.7
10	2.1	1.5	1.3	.9	1.1
Sum	24.3	21.0	14.2	13.3	13.8
Mean	2.43	2.10	1.42	1.33	1.38
$\sigma$	0.61	0.73	0.45	0.71	0.63
SE	0.20	0.24	0.15	0.24	0.21
% of Stand.	32.5%	28.0%	18.9%	17.7%	18.4%

Table 2

Values of *t*, Comparing Mean Magnitude of Errors in Judgments of Equality of the Müller-Lyer Figure at Five Brightness Levels

	Brightness in Foot-Lamberts				
	.005	.01	.05	.10	1.00
.005	—	—	—	—	—
.01	3.62**	—	—	—	—
.05	7.06**	6.42**	—	—	—
.10	7.05**	7.06**	1.34	—	—
1.00	7.45**	6.32**	0.80	0.77	—

\*\* Significant at 1 per cent level.

practice effect and found that the illusion gradually diminished and approached zero. The practice effect held good only for the original position of the figure. Reversal of figures returned the illusion to full strength. The illusion was revived even in the original figure by standing off and looking at it casually as a whole. In our experimental procedure the effect of practice was reduced satisfactorily by the use of units of experimentation containing small number of trials separated by adaptation periods.

Although the experimental design was such as to minimize the effects of errors of habituation and/or expectancy (by giving ascending and descending trials alternately and changing the length of the trials) the analysis of the ascending and descending series at each brightness level shows the ascending series to be greater in all cases than the descending series. The differences were: .58 mm. at .005 foot-lamberts, .74 mm. at .01 foot-lamberts, .96 mm. at .05 foot-lamberts, 1.06 mm. at .10 foot-lamberts, and .92 mm. at 1.00 foot-lamberts. These are errors of habituation.

All subjects commented that judgments were more difficult to make under the two lower brightnesses but they all reported that they thought they did as well under the lower brightnesses as under the higher levels.

### Discussion

The results reported above indicate for this task a critical level of brightness (about .05 lamberts) below which the magnitude of perceptual errors increases significantly from the errors at and above this brightness level.

Above this critical level further increases in brightness up to 1 foot-lambert (and possibly indefinitely), produce no significant increments of performance. It would seem as though once a subject has been given sufficient brightness to perform the task with ease, brightness is no longer a significant variable.

### Experiment II. Absolute Motion Threshold

One of the first questions to be raised with reference to the amounts of the motion that are either just perceptible or just not perceptible, is that of the so-called threshold value. In this study only the lower absolute threshold is to be considered. It is found that when velocity of a stimulus is diminished, there is a critical level of velocity beneath which no motion is perceived. Perception of motion not only depends on: (1) the physical velocity of the moving stimulus, but also on such variables as (2) form and size of the stimulus; (3) presence or absence of fixed reference objects, and their nature; (4) absolute and relative brightness of the stimulus and the background; (5) absolute and relative color of stimulus and background; (6) light or dark adaptation of the eye; (7) monocular and binocular observation; (8) macular or peripheral observation of the stimulus; (9) distance of observation; (10) duration of the observation period; (11) eyes allowed to move or required to fixate; and (12) characteristics of the path of movement. Under usual operational conditions the observation is binocular with macular and/or peripheral fixation, non-limited duration of observation period, and the eyes moving in normal manner.

In the present experiment the variables were treated in the following manner: The relation of physical velocity of the moving stimulus to the absolute and relative brightness of the stimulus and background was measured; absolute and relative color of the stimulus and background, size and form of stimulus, adaptation of the eye, distance of the observation, and characteristics of the path of movement were all controlled.

### Method

*Apparatus.* The general plan of the experimental situation has been described in the preceding study.

The subject was seated in the three-sided booth facing the front wall. In the front wall was a  $2 \times 4$ -inch aperture over which was superimposed a double gradient neutral density filter with the center clear and increasing in density toward the ends. The stimulus grid was presented in this aperture and the double gradient neutral density filter acted to gradually blur the edges so that no sharp reference boundary was evident in the field.

The stimulus grid consisted of high-contrast photographic reproductions of 2 mm. wide alternate black and white lines. The grid was a continuous circular band, 2 inches in height and 18 inches in circumference. It was carried on three rollers placed so as to form a triangle  $7\frac{1}{2}$  inches on the aperture side,  $5\frac{1}{2}$  inches and  $4\frac{1}{2}$  inches on the other two sides. The rollers were  $\frac{3}{4}$  inch on the other two sides. The roller not in the rubber covered dowls. The roller not in the aperture was driven by a 1/60 horsepower General Electric A-C motor, model 5KH13E 19, type K.J., with a friction clutch attachment in conjunction with a 1/100 reduction gear. The actual velocity of the driven shaft was recorded in R.P.M. by a tachometer. The velocity of the driven shaft could be changed continuously and smoothly through a range of .05 mm. per sec. to 2 mm. per sec. A masking motor was employed in conjunction with the apparatus so as to mask any possible noise cues.

The experimenter was seated at a small work table placed against the outside of the middle wall of the booth. The velocity regulating knob was within easy reach and the dial face of the tachometer was directly in front of him. On the table before him were located the motion apparatus described above, a Variac and a voltmeter for control of the subject's lights, a carefully hooded lamp to provide minimal illumination and a data sheet to record velocities and subject remarks.

The light sources and controls were as in the previous experiment.

Data sheets were prepared in advance as in the previous experiment.

**Subjects.** The same ten subjects who served in the preceding study were used in the present experiment. A month or more elapsed between their participation in the two experiments.

**Procedure.** Each subject was allowed to become dark adapted before the illumination was turned on. With .05 foot-lamberts of brightness and the stimulus velocity very low (subliminal) the subject was instructed:

This is an experiment to determine the influence of varying brightnesses of illumination on the moving of strips. You are to look at the center of the screen, look at three or four strips and say "Now" when you see these strips move in a regular fashion across the screen. (Increase the velocity well above threshold.) Now the strips are moving across

Table 3

Showing the Mean Tachometer Readings at the Point of Absolute Motion Threshold and Mean Values for Minutes of Arc/Second at Five Brightness Levels

Subject	Brightness in Foot-Lamberts					
	.005	.01	.05	.1	1.00	10.00
1	1.4	1.4	1.3	1.0	.6	
2	1.2	1.1	1.1	.7	.5	
3	1.4	1.4	1.4	.6	.5	
4	1.5	1.5	1.4	.8	.8	.7
5	1.3	1.3	1.3	.7	.6	
6	1.4	1.4	1.4	.7	.5	.5
7	1.5	1.4	1.4	.8	.8	
8	1.5	1.2	.9	.5	.4	
9	1.6	1.3	1.3	.7	.7	
10	1.4	1.0	1.0	.6	.6	.6
Mean Tach. Read. (RPM)	1.42	1.30	1.25	.71	.60	.60
Mean, in Min. of Arc/Sec.	.40	.36	.35	.20	.17	
$\sigma$ Tach.	0.10	0.15	0.17	0.13	0.13	
SE Tach.	0.03	0.05	0.06	0.04	0.04	

the screen in a regular fashion; say "Now" when you can't see them moving in this regular fashion.

### Results

The data of this experiment consist of judgments of the presence or absence of motion (absolute motion perception thresholds) for ten subjects under five different illumination levels.

The means of ten judgments for each subject are presented in Table 3 which shows for each subject not only the mean velocity in R.P.M. but also in minutes of arc per sec. at each of the five brightness levels.

Inspection of Table 3 and Figure 2 suggests that: (1) the absolute motion perception threshold is markedly lower at the higher brightness levels; (2) there is a sharp change in motion perception performance between .05 and .1 foot-lamberts; and (3) there is relatively little improvement in performance above 0.05 foot-lamberts.

Since our principal concern is with performance as a function of brightness level, a *t* analysis was carried out comparing group performance for each pair of brightness levels.

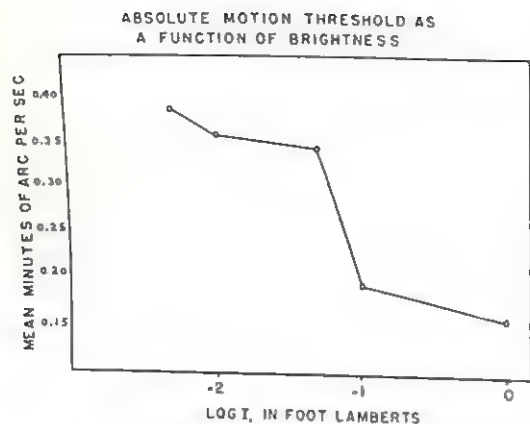


FIG. 2. Visual performance as a function of low photopic brightness levels.

A summary of this analysis is presented in Table 4. From Table 4 it is seen that all differences which cross the .05 foot-lambert level are significant at the 1 per cent level while no difference that does not cross this value is significant at the 1 per cent level. Three other differences are significant at the 5 per cent level: between .005 and .01; .005 and .05; and .1 and 1.0. A supplementary test was made on three of the subjects at 10 foot-lamberts; Table 3 shows that there is no difference between the means at this level and at 1 foot-lambert.

On the basis of the data presented above it seems clear that: (1) motion perception performance increases sharply in the region of .05 foot-lamberts; and (2) relatively little increase in absolute motion threshold results from increasing brightness above this level.

**Practice Effects.** The method employed in the investigation of practice effects was the same as in Experiment I. Each brightness level appeared in each ordinal position an equal number of times. A *t* analysis of the several differences indicated no significant practice effects in this experiment. The analysis of the ascending and descending series at each brightness level shows the ascending series to be greater in all cases than the descending series; these are errors of habituation. These errors of habituation prove to be small and relatively constant for all brightness levels. The differences were: .04 R.P.M. at .005 foot-lamberts, .12 R.P.M. at .01 foot-lamberts, .18 R.P.M. at .05 foot-

lamberts, .10 R.P.M. at .1 foot-lamberts, and .08 R.P.M. at 1.00 foot-lamberts.

The subjects' comments are of interest in this experimental situation. Subjective reports of pulsations in a plane perpendicular to the movement and pulsations in the plane of the movement were frequent. These pulsations developed before motion was apparent, but in decreasing supraliminal motion to no motion the pulsations were usually not reported. The subjects typically believed their performance to be about equally good at all levels of brightness tested.

## Discussion

The results of absolute motion threshold perception reported above indicate a critical level of brightness, between .05 and .1 foot-lamberts, below which subjects' absolute motion threshold is significantly raised. Above this level the absolute motion threshold is a minimum and a further increase in brightness, at least up to 10 foot-lamberts and probably indefinitely, produces no further significant increments of performance. Again it seems as though once a subject has been given just enough brightness to perform this task with ease, brightness is no longer a significant variable.

The absolute motion thresholds found in this experiment ranged from 24 secs. of arc per sec. at .005 foot-lamberts to 10 secs. of arc per sec. at 1 foot-lambert (and with three subjects at 10 foot-lamberts). These results are lower than those usually stated as typical. Generally one to two minutes of arc per sec. is the absolute motion threshold reported.

Table 4

Values of *t*, Comparing Mean Tachometer Reading at Point of Absolute Motion Threshold at Five Brightness Levels

	.005	.01	.05	.1	1.00
.005	—	—	—	—	—
.01	2.44*	—	—	—	—
.05	2.69*	1.63	—	—	—
.10	12.52**	12.94**	10.82**	—	—
1.00	19.72**	14.17**	13.00**	2.70*	—

\* Significant at 5 per cent level.

\*\* Significant at 1 per cent level.

The source quoted is usually Aubert's experiment, but his results should be qualified with, "under short fixation times," for Aubert follows his results with: "... whereas with lower velocities it requires several seconds to detect motion" (16). With unlimited fixation time Munch (39) reported thresholds as low as 34 secs. of arc per sec. and Basler (39) as low as 13 secs. of arc per sec. for foveal fixation under daylight conditions. This compares closely with the present results. Thus, it would seem that under ordinary operational conditions with unlimited fixation time, brightness above .05 foot-lamberts, binocular vision, cone adapted eyes, with a blurred but stationary reference in the visual field, and stimulus size subtending 10 minutes of arc the absolute motion threshold is of the order of 10 sec. of arc per sec. Below a brightness of .05 foot-lamberts the motion threshold increases rapidly up to 24 sec. of arc per sec. at .005 foot-lamberts.

### Experiment III. Depth Perception

The two preceding experiments have presented data concerning performance on a visual illusion and on a motion discrimination task as a function of low photopic brightness levels. The present study is a further extension of these studies to performance on a depth perception task as a function of the same low photopic brightness levels.

It is evident that it is impossible for the retinal image alone to give us a tridimensional perception, for the retinal image is only two dimensional. Stereopsis can best be considered a unification of many visual impressions and among the factors utilized to a greater or lesser extent are: (1) size of retinal image, (2) aerial perspective, (3) mathematical perspective, (4) distribution of lights and shadows, (5) intervening objects, (6) convergence and accommodation, and (7) parallax.

The most important factor contributing towards binocular stereopsis is the phenomenon of binocular parallax. It is the absence of this factor in monocular vision that renders the estimation of depth so difficult, especially with the head fixed.

### Method

*Apparatus.* The experimental situation has been described above.

For this experiment the stimulus field contained three dull white rods, the two outer rods fixed and the center rod movable. The rods were separated by  $\frac{3}{4}$ " and were 2 mm. in diameter. Approximately one inch in the center region of the rods was visible to the subject. The background was a matte black. The movable rod was moved by means of a rack and pinion by the experimenter. Brightness was controlled as in the previous experiments.

*Subjects.* The same ten subjects who served in the previous studies were used in the present study. Two weeks or more elapsed between their participation in this experiment and the preceding one.

*Procedures.* Each subject was allowed to become cone dark adapted (approximately 10 minutes) before the illumination was turned on. The instructions were as follows:

This is an experiment to determine the influence of varying brightnesses of illumination on the performance of depth perception. You are to look at the three white rods. The center one will be moved back and forth; you are to tell me when you see it in the same plane as the other two. Now the center rod is back of the other two. Say "equal" when it appears to be in the same plane as the two fixed rods; now tell me when it appears in front of the other rods. Now the rod is well in front of the other two rods. Say "equal" when it appears to you to be in the same plane as the

Table 5

Showing the Mean Constant Errors in Depth Perception in mm. Under Each of the Five Levels of Brightness

Subjects	Brightness in Foot-Lamberts				
	.005	.01	.05	.1	1.00
1	3.2	2.4	-.9	-.4	-.3
2	3.0	3.9	.7	.9	.3
3	4.7	3.2	.7	.7	-.1
4	1.4	1.2	-.7	-.6	-.3
5	4.3	3.5	.6	.2	1.5
6	3.3	2.1	-1.0	-.2	-.7
7	3.2	1.1	-.4	-.1	-.8
8	2.9	1.7	.9	0	.3
9	7.5	2.8	2.2	1.7	.9
10	1.8	3.5	-.1	-.4	-.4
Mean	3.53	2.54	0.20	0.18	.04
$\sigma$	1.62	0.95	0.52	0.47	0.40
SE <sub>M</sub>	0.54	0.32	0.17	0.16	0.13

others. Now tell me when it appears behind the other two rods.

Five ascending (toward observer) and five descending (away from observer) trials were given before the formal trials were begun; this fore-test served to reduce the practice effect.

On the formal trials each subject was given five ascending and five descending trials at each of five brightness levels. The method used was the traditional method of limits.

The levels of brightness were the same as the preceding studies.

Results

The data of this experiment consist of constant errors, and average errors (non-algebraic), made by the group of ten subjects under the five levels of brightness employed.

The constant error data are presented in Table 5, which shows for each subject the mean of 20 judgments at each brightness level. Mean values for the group are thus based on 200 judgments at each level. The mean values from Table 5 are shown in Figure 3 as a function of log brightness in foot-lamberts.

Inspection of Table 5 and of Figure 3 indicates that: (1) performance is markedly more accurate at the higher brightness levels; (2) there is a relatively sharp change in performance between .01 and .05 foot-lamberts; and (3) the mean constant errors are all positive in direction, i.e., at the judgment of equality the variable is in all cases in front of the two fixed rods. A *t* analysis was carried out comparing group performance for each pair of brightness levels and a summary of

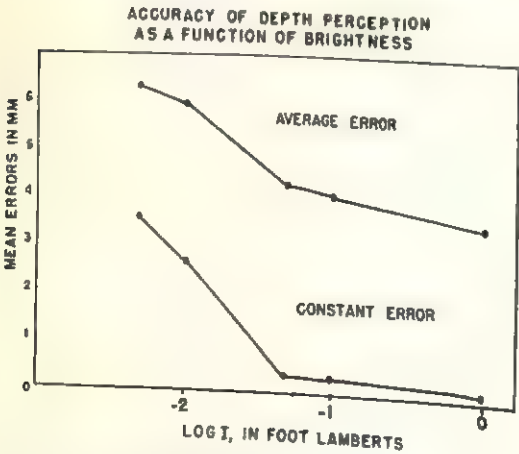


FIG. 3. Visual performance as a function of low photopic brightness levels.

Table 6

Values of *t*, Comparing Constant Errors in Depth Perception at Five Brightness Levels

	Brightness in Foot-Lamberts				
	.005	.01	.05	.1	1.00
.005	—	—	—	—	—
.01	1.81	—	—	—	—
.05	7.73**	6.88**	—	—	—
.1	9.31**	8.14**	1.33	—	—
1.00	7.76**	8.93**	0.76	0.66	—

\*\* Significant at 1 per cent level.

this analysis is presented in Table 6. From this table it is seen that all differences which cross the .01 foot-lambert level are significant at the 1 per cent level, while no difference that does not cross this value is significant at the 5 per cent level.

The data for average errors are presented in Table 7 and the *t* analysis in Table 8. Figure 3 presents the results graphically. The results are on the whole the same for the two kinds of error analysis. The only noticeable difference is that the average errors continue to decrease above the critical brightness level while the constant error values change very little above this point. Although (see Table 8) the differences between .05 and .1 and be-

Table 7

Showing the Mean Average Errors in Depth Perception in mm. Under Each of the Five Levels of Brightness

Subjects	Brightness in Foot-Lamberts				
	.005	.01	.05	.1	1.00
1	5.5	5.9	4.4	4.1	4.1
2	4.8	6.0	2.9	2.3	2.1
3	5.2	4.1	3.3	2.7	2.4
4	6.0	5.5	3.4	3.3	3.0
5	6.5	5.4	3.0	3.1	3.2
6	6.1	5.2	3.6	4.4	3.6
7	5.5	6.3	4.6	4.4	3.7
8	6.6	6.4	5.9	4.4	4.4
9	10.4	7.1	4.6	5.1	3.4
10	6.6	6.6	5.6	5.6	5.2
Mean	6.3	5.9	4.1	3.9	3.5
σ	1.5	0.8	1.0	1.0	0.9
SE <sub>M</sub>	0.5	0.3	0.3	0.3	0.3

Table 8

Values of *t*, Comparing Average Errors in Depth Perception at Five Brightness Levels

	Brightness in Foot-Lamberts				
	.005	.01	.05	.1	1.00
.005	—	—	—	—	—
.01	0.98	—	—	—	—
.05	4.53**	6.27**	—	—	—
.1	6.15**	12.50**	1.00	—	—
1.00	5.44**	11.71**	3.40**	2.27	—

\*\* Significant at 1 per cent level.

tween .1 and 1.00 foot-lamberts are not significant, the difference between .05 and 1.00 is significant at 1 per cent level.

Analysis of the above data shows that: (1) Accuracy of depth perception performance decreases sharply below .05 foot-lamberts and (2) little increase in accuracy results from increases in brightness above this level.

The results of this experiment, expressed in terms of angle of binocular parallax, are shown in Table 9.

**Practice Effects.** Practice or "warm-up" effects were analyzed as in the other experiments. A *t* test analysis of the several differences indicate no evidence of a practice effect for the experiment.

The subjects' comments were of interest in that seven of the ten subjects made some reference to the apparent decrease in triangularity. All these seven subjects stated in various ways that the center movable rod was equal when the apparent triangularity was zero and that the fixed rods were used as the base standard.

### Discussion

The results of the constant and variable error reported above clearly indicate that for depth judgments there is a critical level of brightness between .01 and .05 foot-lamberts. Depth perception is relatively difficult below this level. Above this value the task becomes suddenly easier and both constant and variable errors decrease markedly. Further increases in brightness—at least up to 1.0 foot-lamberts and probably indefinitely—produce no further increments of performance.

It is interesting to note that although Muel-

ler and Lloyd (20) state that stereoscopic acuity decreases in a regular fashion as intensity decreases, if one plots their results so that the actual data points are connected and the curve not smoothed a sharp break appears at approximately the same brightness level as in the present experiment. It is of interest also to note that the trend of their results would be highly similar to that of the present experiment, with performance leveling off above approximately .05 millilamberts.

The absolute binocular parallax values found in this experiment for some subjects are lower than reported by the earlier experimenters. Bourdon's (2) 5 seconds was the lowest difference reported. Six of our ten subjects had five seconds or less at brightness level above the critical point of .05 foot-lamberts. The variability of these at the adequate performance brightness levels was great.

These results add another item to our knowledge of critical visual performance levels as a function of low photopic brightness levels. Depth perception as well as dial reading performance, illusion errors and absolute motion thresholds appear to have approximately the same critical level of brightness, above which performance is adequate and an increase in brightness has relatively little effect and below which performance is relatively poor.

### Experiment IV. Addition Task

A simple quantifiable mental task was investigated to add to the picture of visual performance at low photopic brightness levels. Reading tasks are the first to come to mind, but two serious objections are inherent in the use of reading material. First, only time

Table 9

Showing the Mean Constant Errors and Average Errors as Angles of Binocular Parallax

Brightness, in Foot-Lamberts	Constant Error (Sec. of Arc)	Average Error (Sec. of Arc)
0.005	90	160
0.01	60	155
0.05	5	105
0.1	5	100
1.0	<5	90

scores can be made with any reliability, and second, at low photopic levels the material would have to be made abnormally large in order to avoid the factor of visual acuity as a limiting variable. Two of the most productive investigators in this field, Tinker and Luckiesh, have reported numerous studies, of which the two following are typical. Tinker (34), investigating reading of 10 point type at illuminations ranging from 0.1 to 53 foot candles (accuracy held constant), found speed of reading to increase rapidly from 0.1 to 3.0 foot candles and no change in speed of reading between 3.0 to 53.0 foot candles. No change in accuracy of reading was reported. Luckiesh and Moss (18) in correlating illumination intensity and nervous muscular tension resulting from reading 12 point type (large type) found the critical intensity level to be somewhere between 1 and 10 foot candles.

Addition tasks have been used by many investigators as an active mental task. Thorndike (31) used addition scores of time and accuracy to investigate practice. Davis (8) used addition tasks to investigate the effect of noise on mental work. Rounds, Schubert, and Poffenberger (26) used addition tasks to investigate the effects of practice upon the metabolic cost of mental work. Freeman (13) employed mental arithmetic as a mental task and measured it as a function of spread of neuromuscular activity. Addition tasks have also been used by many other investigators as a mental task. Atkins (1) used arithmetic problems of the cancellation type as a mental task and measured it as a function of illumination. His range was from 9.6 to 118 foot candles. Performance measured by achievement was identical at all levels of brightness.

In the present study performance as measured by both speed and accuracy in addition problems was investigated as a function of low photopic brightness levels.

### Method

*Apparatus.* The basic experimental situation has been described in the preceding experiments. In the stimulus position for this experiment was a stimulus card carrier which slid in horizontally placed brass tracks. It was double (11 × 14 inches) so that as one stimulus object was slid out of the subject's view, another stimulus ob-

ject came immediately into view. Micro-switches at each end of the track were arranged so that illumination on the stimulus object went off as the carrier was moved from one position and came on as it reached the other position. In this way the shift from one stimulus object to another was accomplished rapidly in a short interval of darkness and did not require the subject to make any shift in visual orientation. Thus, the subject was kept steadily at the chosen level of illumination throughout a series of readings, except for an instant of darkness between the presentation of stimulus objects.

*Materials.* The stimulus objects, generously made available to this project by Dr. Mason Crook and Sam McLaughlin of Tufts College Air Forces project, consisted of high-contrast photographic reproductions of 10 point monotype numbers arranged into 100 items per chart, each item consisting of a 3-digit problem and its 2-digit sum arranged horizontally. There were five items per group, four groups per column, and five columns per chart. Each chart was reproduced so as to have white figures on a black background and size was increased two-fold so that each digit had an over-all height of 3/16 inches. The problems were selected with predetermined specifications as to random numbers, repetitions, zeros, sums, et cetera.

*Subjects.* The same subjects of the previous experiments participated in this experiment. A week or more passed between this experiment and the preceding experiment.

*Procedure.* Each subject was given 10 charts at a brightness of 10 foot-lamberts on the day preceding the day of the formal trials; this pretest served to reduce practice effects.

On the formal trials each subject did two charts (200 problems) at each of five brightness levels. While each subject was becoming comfortable adapted the following instructions were read:

This is an experiment to determine the influence of varying brightnesses of illumination on the performance in simple addition problems. There are 5 columns of numbers, each column separated after 5 problems. The first three numbers are to be added up and if their total equals the fourth number, you are to say "right"; if they add up to some other number you are to say "wrong." You will say "space" after each 5 problems to indicate the spaces on the charts; you are to say "new column" when you start each new column.

When I say "ready" the lights will go off (subject is preadapted to brightness level to be used by looking at a matte black chart illuminated with this brightness) and in a moment they will come on again. When the lights come on, you are to add the numbers as directed and say "right" or "wrong" to each problem; remember to say "space" after each

Table 10

Showing the Total Number of Problems in Error in 200 Addition Problems at Each of Five Brightness Levels

Note: The task was impossible for all subjects at .005 foot-lamberts.

Subjects	Brightness in Foot-Lamberts				
	.008	.01	.05	.1	1.00
1	30	29	1	2	1
2	46	36	0	1	0
3	48	21	1	0	1
4	70	54	5	1	2
5	12	6	1	1	2
6	39	26	2	5	1
7	50	34	2	3	3
8	53	55	16	4	5
9	20	23	6	2	2
10	37	24	3	1	3
Mean	40.5	30.8	3.7	2.0	2.0
$\sigma$	15.84	14.16	4.38	1.48	1.34
SE <sub>M</sub>	5.28	4.72	1.46	.49	.45

set of 5 problems and remember to say "new column" when going from one column to the next one. You are to add down the first column on the left and proceed to the right. Add as *rapidly* and as *accurately* as you can.

The brightness levels were originally the same as the preceding experiments, but all subjects found the task impossible under the .005 foot-lambert level. The lower level was raised to .008 foot-lamberts, at which brightness better than chance results were obtained. The levels finally used were .008, .01, .05, .1, and 1.00 foot-lamberts. A piece of unexposed but developed paper from the same stock as the stimulus charts was used as an object for this standardization.

Controls to balance out practice and fatigue effects were the same as in the previous experiments.

Each subject reported for two sessions on consecutive days. At the first session subjects were given the fore-test and on the second day the formal trials were given. Subjects were given no knowledge of results during the entire experiment.

### Results

The data of this experiment consist of error scores and time scores made by the same group of 10 subjects under 5 brightness levels. It is interesting to note that at the .005 foot-lambert level all subjects felt the task to be impossible and refused to attempt it, reporting that the task would be guesswork and

feared injury to eyes. The lowest level was raised to .008 foot-lamberts where all subjects did better than chance, although still reporting difficulty at this low level.

**Errors.** The principal analysis of errors is in terms of error frequency, i.e., the number of problems in error in 200 problems. These data are presented in Table 10, which shows the total number of problems in error in 200 addition problems at each of the five brightness levels. Mean values of the group are thus based on a total of 2,000 readings at each level. The mean values from Table 10 are also shown as the accuracy curve of Figure 4.

From Table 10 and Figure 4 inspection indicates that: (1) performance is markedly more accurate at the higher brightness levels; (2) there is rapid improvement in performance up to .05 foot-lamberts; and (3) little or no improvement above .05 foot-lamberts.

A *t* analysis was carried out comparing group performance for each pair of brightness levels. A summary of this analysis is presented in Table 11. From this table it is seen that all differences between .05 foot-lamberts and lower brightness levels are highly significant (1 per cent level) and all differences between .05 foot-lamberts and higher brightness levels are not significant at the 5 per cent level.

From this analysis it seems clear that: (1) accuracy of performance in doing addition problems increases sharply in the region between .01 and .05 foot-lamberts; and (2) little or no increase in accuracy results from increases in brightness above this level.

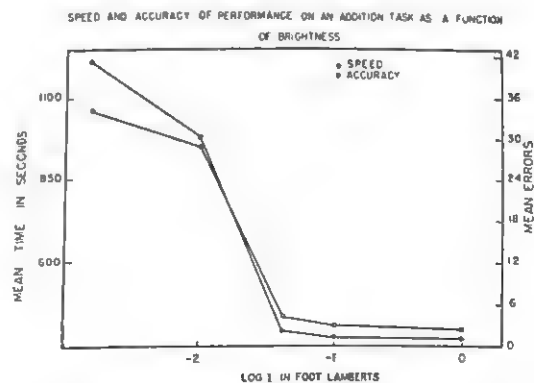


FIG. 4. Visual performance as a function of low photopic brightness levels.

Table 11

Values of  $t$ , Comparing Mean Number of Errors in 200 Addition Problems at Five Brightness Levels

	Brightness in Foot-Lamberts				
	.008	.01	.05	.1	1.00
.008	—	—	—	—	—
.01	3.28**	—	—	—	—
.05	7.22**	6.84**	—	—	—
.1	7.13**	6.26**	1.26	—	—
1.00	7.40**	6.40**	1.48	0.00	—

\*\* Significant at the 1 per cent level.

**Time.** Data on the speed of performance in doing addition problems at the brightness levels studied consist of the time in seconds required to do 200 addition problems at each brightness level. Each subject did two charts of 100 problems each at each brightness level. Table 12 presents the total time required to do 200 problems for each subject at each brightness level, and also the group means. The mean values from Table 12 are also shown as the speed of response curve of Figure 4.

Figure 4 shows that the curve for time scores has the same general shape as the error curve. There appears to be a sharp increase in performance between .01 and .05 foot-lam-

Table 12

Showing the Total Time in Seconds Required to Do Addition Task for 200 Problems at Each of Five Brightness Levels

Subjects	Brightness in Foot-Lamberts				
	.008	.01	.05	.1	1.00
1	1569	1422	687	623	634
2	1189	903	348	358	301
3	1019	839	419	416	338
4	964	1099	420	359	298
5	622	571	414	415	406
6	1083	1007	367	389	356
7	1394	1046	502	446	426
8	770	865	414	370	332
9	727	694	448	424	411
10	1344	1105	476	426	418
Mean	1068.1	955.1	449.5	422.6	392.0
$\sigma$	294.2	226.3	90.9	72.7	92.6
SE <sub>M</sub>	98.1	75.4	30.3	24.2	30.8

Table 13

Values of  $t$ , Comparing Mean Time in Seconds to Do 200 Addition Problems at Five Brightness Levels

	Brightness in Foot-Lamberts				
	.008	.01	.05	.1	1.00
.008	—	—	—	—	—
.01	2.56*	—	—	—	—
.05	8.48**	8.50**	—	—	—
.1	7.51**	5.76**	2.64*	—	—
1.00	8.08**	8.70**	5.28**	3.45**	—

\* Significant at 5 per cent level.

\*\* Significant at 1 per cent level.

berts. The  $t$  analysis in Table 13 shows that all differences of means are significant at the 5 per cent or 1 per cent level. The difference between .008 and .01 foot-lamberts is significant at the 5 per cent level, and the difference between .05 and .1 foot-lamberts is significant at the 5 per cent level; all the other differences are significant at the 1 per cent level. Thus, it can be seen that even though the general results found for time agree with those for errors, there is evidence to support the statement that performance as measured by speed of doing addition tasks increases significantly with increased brightness up to 1.0 foot-lamberts.

**Practice Effects.** It will be recalled that each subject did 1,000 addition problems before formal trials were begun, in order to reduce practice effects. As in the previous experiments, in order to determine whether practice effects were playing a significant role, the error and time scores were tabulated for each subject in terms of first brightness level tested, second level tested, etc. The  $t$  test analysis showed no evidence of practice effects.

### Discussion

The results of the error scores indicate clearly that there is a critical level of brightness below which subjects find it difficult to perform this addition task. Above this level further increases in brightness, at least up to 1 foot-lambert and very probably indefinitely, produce no significant increments of performance. These results agree completely with those from the preceding experiments.

The results of the time scores, although comparable to the error scores in general trend, indicate that speed of performance increases as brightness increases, at least up to 1.00 foot-lambert. The increase shown by the significance of the several differences is not a steady increase but has its greatest rate between .01 and .05 foot-lamberts.

These findings agree with the findings of the three preceding experiments in showing that performance in an active mental task as a function of brightness shows a critical value at the same general level of brightness. This critical brightness level is between .01 and .05 foot-lamberts which is considerably lower than the value of 3 foot-lamberts usually reported as critical for reading performance. In such studies speed of reading has usually been the criterion of performance. It has already been noted that visual acuity may complicate the picture of reading at low photopic levels and it has been noted that in some studies accuracy did not change over a range of .1 to 53 foot-candles when words were large enough to be read. The time scores in this study indicate an increase of performance up to 1 foot-lambert, which was the highest value used, but this increase was a differential increase with slower rates between .05 to .1 to 1.0 foot-lamberts, and a sharp increase between .01 and .05 foot-lamberts. Error scores in this study decreased rapidly up to .05 foot-lamberts and then leveled off, showing no significant decrease for the higher levels of brightness.

#### Over-all Discussion

The present experiments have been concerned with visual perceptual performance as a function of low photopic brightness levels. The functions found, like those of earlier dial reading studies from this laboratory, differ markedly from the functions which have frequently been reported from studies of the effects of brightness on visual acuity and foveal flicker fusion frequency. These latter functions have been found by numerous investigators to increase steadily in a manner proportional to the logarithm of the stimulus intensity.

In contrast the results from the four ex-

periments reported here, as well as from the preceding dial reading experiments, indicate that performance improves as stimulus intensity increases only up to a certain point (approximately 0.05 foot-lamberts, depending somewhat on the specific task employed). Beyond this point increases in stimulus intensity are relatively unimportant in these experiments, the increments in performance being small and non-significant.

These findings raise a number of interesting theoretical questions with respect to the physiological mechanisms and inter-relations which may be hypothesized to explain the present results. A detailed account of hypotheses which might account for these findings, and especially of a possible rod-cone facilitation and inhibition relationship, will not be presented here. Such an account has been developed elsewhere in some detail.<sup>1</sup>

In passing it may be of interest to note that a somewhat analogous situation is found in the field of audition. When per cent word articulation is plotted against stimulus intensity, there results a performance curve with sharp increase in the 10 db. region and a leveling off at about 20 db. (24). Myers and Harris (21) investigating the emergence of a tonal sensation with frequencies from 500 to 14,000 cps. found a "zone of detectability" (intensity area between a 50 per cent detection threshold and a 50 per cent pure tone threshold) between 2 to 4 db., independent of frequency. In their experiment with frequency matching, performance improved with increase in intensity only up to the level of 10 db. It would seem from these studies that in audition, as well as in visual tasks, there is a critical sensation level below which performance is increasingly poor and above which increases in stimulus intensity do not increase appreciably the subject's performance.

From a practical standpoint the results of the present experiments suggest certain minimum values for adequate performance of visual tasks. From the present study and other available sources we can summarize for

<sup>1</sup> In the original from which this report was rewritten, on file as a doctoral dissertation in the University of Rochester Library.

a variety of visual tasks critical brightness levels, below which performance is impaired:

Müller-Lyer illusion	between .01 — .05 F.L.
Depth perception	between .01 — .5 F.L.
Motion discrimination	between .05 — .1 F.L.
Addition task	between .01 — .05 F.L.
Dial reading	approx. .02 F.L.

Critical fusion frequency (cone) .05 F.L.

Span of apprehension (.032 sec. — 1 sec.) .1 — .05 F.L.

Panel indicator lights between .01 — .1 e.f.c.

Form silhouettes above .003 e.f.c.

In view of the above findings it might seem advisable to consider .05 to .1 foot-lamberts (equals .1 e.f.c.), which is one of the highest values given, to be the limiting values to be used in practical situations. This indicates that in situations where the maximum quality and quantity of performance is required with the minimum brightness a value of approximately .05 to .1 foot-lamberts should be employed. Lighting of airplane cockpits, automotive and rail operator compartments and other situations which require good visual performance in the operator's compartment plus adequate dark adaptation permitting for good form discrimination are situations to which this finding is relevant.

Somewhat aside from the present data but as a rather interesting extension is the proposed use of a flood-type light yielding this critical brightness level in the operator's compartment (so situated as not to give reflections from the windshield, etc.). This should serve to raise the adaptation level of the eyes to the critical level where form and silhouette discrimination is adequate. On-coming headlights or disturbing flashes of various types should have less "blinding" or "dazzle" effect because the adaptation change of the eyes would be less than that which is now required (from dark or near dark adaptation to bright on-coming lights or to flashes of lights). Since the visual performance that is needed by the operator is one of form, silhouette, depth perception, motion acuity and minimization of illusion, etc., his visual performance outside the compartment should also improve or at least not be impaired. An

emphasis on the physiological cause of "dazzle" rather than the changing of the physical and optical constituents of headlights, windshields, gun flashes, etc., may be a more fruitful approach to the problem.

### Summary

A systematic investigation of performance in visual tasks as a function of low photopic brightness levels was attempted. Four types of visual tasks were investigated: judgment of magnitude of an illusion, absolute threshold for motion, depth perception and a simple addition task. All tasks were investigated under five brightness levels in the range of .005 foot-lamberts to 1.00 foot-lamberts. In each of the experiments, critical brightness levels were found below which performance was increasingly poor. Increased brightness above the critical level improved performance relatively little or not at all. The critical level for motion threshold was .1 foot-lamberts; for the other tasks approximately .05 foot-lamberts. It was suggested that for maximum performance on visual tasks, with minimum brightness, illumination should be adjusted to yield brightness values of .05 to .1 foot-lamberts.

Received May 28, 1953.

Early publication.

### References

1. Atkins, E. W. The efficiency of the eye under different intensities of illumination. *J. comp. Psychol.*, 1927, 1, 1-37.
2. Bourdon, B. *La perception visuelle de l'espace*. Paris: Libraire C. Reinwald, Schleicher Frères, Editeurs, 1902, Pp. 432.
3. Brown, J. F. The visual perception of velocity. The thresholds for visual movement. *Psychol. Forsch.*, 1931, 14, 190-232, 249-268.
4. Brown, J. F. and Mize, R. H. On the effect of field structures on differential sensitivity. *Psychol. Forsch.*, 1931, 15, 355-372.
5. Cobb, P. W. Some experiments on speed of vision. *Trans. Illum. Engng. Soc.*, 1924, 19, 150-175.
6. Cobb, P. W. The relation between field brightness and the speed of retinal impressions. *J. exp. Psychol.*, 1925, 8, 77-108.
7. Craik, K. J. W. Legibility of different coloured instrument markings and illuminated signs at low illuminations. *Gt. Brit. Ministry. F.P.R.C.* 415, 15 January 1942, Pp. 4.

8. Davis, R. C. Modification of the galvanic reflex by daily repetition of a stimulus. *J. exp. Psychol.*, 1934, 17, 504-535.
9. Ferguson, H. H. and McKellar, T. P. H. The influence of chromatic light stimulation on the subsequent rate of perception under conditions of low illumination. *Brit. J. Psychol.*, 1943-44, 34, 81-88.
10. Ferree, C. E. and Rand, G. Intensity of light and speed of vision studied with special reference to industrial situations. Part I. *Trans. Illum. Engng. Soc.*, 1922, 17, 69-102.
11. Ferree, C. E. and Rand, G. The effect of intensity of illumination on the near point of vision and a comparison of the effect for presbyopic and non-presbyopic eyes. *Trans. Illum. Engng. Soc.*, 1933, 28, 590-611.
12. Ferree, C. E. and Rand, G. The effect of increase of intensity of light on the visual acuity of presbyopic and non-presbyopic eyes. *Trans. Illum. Engng. Soc.*, 1934, 29, 293-313.
13. Freeman, C. L. The speed of neuro-muscular activity during mental work. *J. gen. Psychol.*, 1931, 5, 479-494.
14. Graham, C. H. and Hunter, W. S. Thresholds of illumination for visual discrimination of direction of movement and for the discrimination of discreteness. *J. gen. Psychol.*, 1931, 5, 178-190.
15. Hartline, H. K. *Relative merits of lights of different wave length in aircraft cockpit illumination.* U.S.N.R.C.-C.A.M. Report No. 10, June 1941, Pp. 1.
16. von Helmholtz, H. L. F. *Treatise on physiological optics.* Translated by J. P. C. Southall. *J. Opt. Soc. Amer.*, 1925, 3, x + 688.
17. Luckiesh, M. and Moss, F. K. *Seeing: A partnership of lighting and vision.* Baltimore: Williams and Wilkins, 1931, Pp. 248.
18. Luckiesh, M. and Moss, F. K. A correlation between illumination intensity and nervous muscular tension resulting from visual effort. *J. exp. Psychol.*, 1933, 16, 540-555.
19. Luckiesh, M. and Moss, F. K. *Reading as a visual task.* New York: D. Van Nostrand Co., 1942, 315-335.
20. Mueller, C. G. and Lloyd, V. V. Stereoscopic acuity for various levels of illumination. *Proc. Nat. Acad. Sci.*, 1948, 34, 223-227.
21. Myers, C. K. and Harris, J. D. *The emergence of a tonal sensation.* Medical Research Dept., U. S. Submarine Base, New London, 31 March, 1938, Pp. 10.
22. McFarland, R. A. *Human factors in air transport design.* New York: McGraw-Hill, 1946, 433-486.
23. McFarland, R. A., Knehr, C. A., and Berens, C. Metabolism and pulse rate as related to reading under high and low levels of illumination. *J. exp. Psychol.*, 1939, 25, 65-75.
24. Office of Scientific Research and Development. Summary technical report of division 17. N.D.R.C. Vol. III. Transmission and reception of sounds under combat conditions. Washington, 1946, 69-108.
25. Rock, M. L. *Annotated bibliography on visual performance at low photopic illumination levels.* U.S.A.F. Air Materiel Command. AF Technical Report 6013, November 1950. Pp. 31.
26. Rounds, G., Schubert, H., and Poffenberger, A. T. Effects of practice upon the metabolic cost of mental work. *J. gen. Psychol.*, 1932, 7, 65-79.
27. Senders, V. L. The physiological basis of visual acuity. *Psychol. Bull.*, 1948, 45, 465-490.
28. Sheard, C. The effects of intensity of illumination on presbyopia, accommodation and convergence. *Amer. J. Opt.*, 1936, 13, 241-254.
29. Spragg, S. D. S. and Rock, M. L. *Dial reading performance as related to illumination variables. I. Intensity.* U.S.A.F., Air Materiel Command, Memorandum Report MCREXD-694-21. 1 October 1948. Pp. 32.
30. Spragg, S. D. S. and Rock, M. L. *Dial reading performance as related to illumination variables. II. Spectral distribution.* U.S.A.F., Air Materiel Command, Memorandum Report MCREXD-694-21A. 1 December, 1948, Pp. 23.
31. Thorndike, E. L. Practice in the case of addition. *Amer. J. Psychol.*, 1910, 21, 483-486.
32. Tinker, M. A. Illumination and the hygiene of reading. *J. educ. Psychol.*, 1934, 35, 669-680.
33. Tinker, M. A. Illumination intensities for reading. *Amer. J. Ophthal.*, 1935, 18, 1036-1038.
34. Tinker, M. A. Illumination standards for effective and comfortable vision. *J. consult. Psychol.*, 1939, 3, 11-20.
35. Tinker, M. A. The effect of illumination intensities upon fatigue in reading. *J. educ. Psychol.*, 1939, 30, 561-571.
36. Tinker, M. A. The effect of adaptation upon visual efficiency in illumination studies. *Amer. J. Optom.*, 1942, 19, 143-151.
37. Tinker, M. A. Criteria for determining the readability of type face. *J. educ. Psychol.*, 1944, 35, 385-396.
38. Tinker, M. A. Effect of visual adaptation upon intensity of illumination preferred for reading with direct light. *J. appl. Psychol.*, 1945, 29, 471-476.
39. Titchener, E. B. *Experimental psychology.* Vol. I. New York: The Macmillan Co., 1901, 309, 313, 321-327.
40. Weston, H. C. and Taylor, A. K. *The relation between illumination and efficiency in fine work. (Typesetting by hand.)* London: H. M. Stationery Office, 1933, Pp. 24.
41. White, L. R., Britten, R. H., Ives, J. E., and Thompson, L. B. Studies in illumination. II. Relation of illumination to ocular efficiency and ocular fatigue among the letter separators in the Chicago Post Office. *Pub. Hlth. Bull.*, No. 181, Wash. Gov. Printing Office, 1929, Pp. 58.
42. Woodworth, R. S. *Experimental psychology.* New York: Holt and Co., 1938, 647.

## Applied Psychology in Action

### Evaluating Supervisory Training at the Job Performance Level

Theodore R. Lindbom

*Personnel Department, Midland Cooperative Wholesale, Minneapolis, Minn.*

Evaluation of college and university courses, of practical necessity, ordinarily ends with the semester-end examination. The assumption is made, rightly or wrongly, that performance in the examination is correlated with performance in future situations where course content can and should be applied. This is the report of the results of an attempt to evaluate, beyond the classroom level, the performance of a group of University of Minnesota General Extension Division students who had taken a course in supervision. The course is a discussion type course in human relations for supervisors, with emphasis on the recognition of individual differences and the "human element" in supervision, which runs for a semester and consists of 16 evening meetings each 1½ hrs. in length. Practically all students were employed full time with about two-thirds in supervisory capacities. The group studied were students during the spring and fall semesters of 1950 and the spring semester of 1951 in 5 different sections of the class all taught by the writer. This evaluation was made in addition to traditional classroom examinations and test-retest with the standard test, "How Supervise?"

A mailed questionnaire, sent in March, 1952 with one follow-up, produced 66 returns from 129 students. Of these 66, 41 were from persons in supervisory jobs. The analysis of these 41 returns is presented here.

The 2 major questions asked were concerned with behavior changes at two levels: (1) changed behavior of the supervisor on-the-job; and (2) changed behavior of the people he supervised resulting from changed methods of supervision.

In answer to the question, "Is there anything that you are now doing differently—as a foreman or supervisor—because of your experience in these discussions?" 63% answered yes, 27% no, and 10% did not answer or

were undecided. Typical of the comments in answer to this question were:

"More consideration of the employees' problems."

"When employees are by-passed for upgrading, an explanation is given them."

"I am spending more time determining the facts when grievances arise."

"Realizing the personality differences in people and using that in dealing with people."

In answer to the question, "Is there anything about the people you supervise now that is different from the way they were before—anything that has resulted from changes in your operations due to taking 'Elements of Supervision'?" 44% answered yes, 29% no, and 27% did not answer or were undecided. Typical of the comments made when respondents were asked to describe these changes in the people they supervised were:

"Morale is better than in other divisions, not to speak of increased efficiency."

"They show more of an attitude of working 'with' rather than 'for.'"

"My employees come to me for help in almost any type of problem."

"Lower costs of operation due to willingness to cooperate with their foreman and themselves."

Although results indicate that the course was successful, at least to some degree, the study design permits neither definite conclusions nor generalization. The group to begin with was a highly selected one, and the percent return of questionnaires is low enough to allow an additional selection factor to be operating. Time between completion of the course and filling out the questionnaire ranged from 10 to 22 months. Conscious or unconscious misrepresentation of facts by respondents is also a possible factor.

Because of these limitations, the study is not reported for its specific findings or for

generalization. Instead, it is presented as an illustration of an easy-to-make evaluation at a level beyond the traditional classroom examination which appears to measure more di-

rectly the kind of behavior change which the course was intended to bring about—the on-the-job behavior of the supervisor and the people he supervises.

THE JOURNAL OF APPLIED PSYCHOLOGY  
Vol. 37, No. 5, 1953

## Criterion Rationale for a Personnel Research Program

Theodore R. Vallance, Albert S. Glickman, and George J. Suci

*American Institute for Research*<sup>1</sup>

The authors have been engaged in setting up and putting into operation a program for personnel research with naval officers. In any activity of this kind, it is vital to develop a rational framework or "research constitution" for the program. Since it appears that the framework we have developed contains many elements which have general applicability in the planning of personnel research in other educational, industrial, or military settings, the substance of our criterion rationale is presented here.

### Ultimate and Intermediate Criteria

When initiating a program of personnel research we must first ask: What is the nature of the criterion? Until the criterion is defined, assessment of the program, or any of its parts, is not possible. The answer to the question is reflected most ultimately in terms of furthering the objectives of the organization. For the Navy the long-run goals are:

- "1. To defend and support the Constitution of the United States against all enemies.
2. To maintain, by timely and effective military action, the security of the United States, its possessions and areas vital to its interest.
3. To uphold and advance national policies of the United States.
4. To safeguard the internal security of the United States."<sup>2</sup>

The "success" of any organization is meas-

ured by the extent to which its objectives are achieved. However, the fulfillment of "ultimate" organizational aims can seldom, if ever, be directly assessed. Other outcomes, less remote, more susceptible to measurement, must normally be used to evaluate day-to-day operations. Historical hindsight and logical analysis are the usual standards for assuming correlation between the ultimate criterion and subordinate "intermediate" criteria. Contributions to success are then measured at many levels presumed to be correlated with the more ultimate criterion.

The practical research problem, then, is to determine the highest organizational level at which quantifiable measures considered to be reflections of personnel behaviors are possible. In our naval model, ships or command units ashore comparable in size, complexity, and autonomy, represent the organizational entities upon whose efficiency the performance of a given crew member may be expected to have a measurable effect. Consequently, it was taken as a basic assumption that for the measurement of success of individuals, the performance of ships or comparable shore units represents the highest level at which meaningful and practical quantitative criterion measurement can be established. As such these performances also comprise the basis for determining the validity and practical meaningfulness of subordinate criteria—for departments, divisions, smaller groups, and individuals.

### Criteria of Individual Success

Several questions which arise during the process of evaluating naval personnel are be-

<sup>1</sup> This work was done by the Officer Personnel Research Project, at the U. S. Naval Schools Command, Newport, R. I. under contract Nonr 890(00) between the American Institute for Research and the Office of Naval Research.

<sup>2</sup> Key West Agreement, 1948.

lieved to have bearing in programs of evaluation for other kinds of administrators, and executives. These are discussed below.

What is "successful performance" for the individual? At what rank, or after what time on the job, can or should evaluation be made?

The possibility must be recognized that not all officers who are competent at one level of rank or responsibility will be equally competent at the next or other higher levels. (It has not been shown that a good ensign necessarily makes a good admiral.) The question then arises: Is "success" composed of the same factors at each level? If *not*, we are confronted with constantly shifting criteria and must choose the intermediate criterion level that is desirable for a particular class of officers on logical or empirical grounds.

As the correlation between criteria for the several ranks decreases, the risk increases that the selection variables validated against lower-level rank criteria will be unrelated to higher-level rank criteria, or indeed may be negatively related to them.

We are thus confronted with the question: When are we to define "success" as having been achieved? Is it at Officer Candidate School, or when an ensign, or when a commander, or when a chief-of-staff, or is performance in the next superior rank an adequate criterion against which to evaluate performance at immediately subordinate ranks?

We are also confronted with the related problem of deciding what criterion levels to choose for assessing the effectiveness of training. That is, by what performance standards, at what level of responsibility, should the adequacy of training at Officer Candidate School and elsewhere be judged? Should performance right after schooling be taken as the measure of training effectiveness, or should evaluations be made after some specified time has elapsed?

Likewise, with regard to selection, assignment, promotion, retirement, and command, there arises the question of where to look for appropriate standards.

Competency in an executive hierarchy is usually considered to be highly correlated with rank. Ideally, rank and competence in rele-

vant areas should be perfectly correlated. To the extent that the correlation is less than 1.00, room exists for improvement of techniques for evaluating training and duty performance, and for assignment to jobs.

It must be recognized that much of the preceding involves policy decisions at a high level and consists of questions which cannot be answered by a research unit. Lack of such policy decisions leaves the research goals in doubt, leads to confused direction, and lowers the utility of research products.

### Criterion Methodology

The relative status of a variable as a predictor or a criterion is in many cases simply dependent upon the chronological sequence in which variables may be organized. Each intermediate performance criterion presumably should be correlated with a more ultimate criterion and hence serve as a predictor of it. As demanded by exigencies, many performance measures can be considered either criteria or predictors.

Comparability of measures of performance is a basic requirement if such measures are to be used effectively. If success is determined at all ranks and in all duties by the same factors, with differences from rank to rank representing only variations in degree rather than kind of factors involved, then the approach to evaluation is relatively simple. Although this is not likely to be true in most cases, it involves primarily attempts to increase reliability of measures of the factors demonstrated to possess the highest validity as criteria.

Questions of the comparability of criterion measures then crop up with respect to all aspects of criterion measurement and we are faced with the problem of how to render criterion measures equivalent in consideration of differences of rank, raters, ship types, duty, hazard, kind of subordinates, and other situational factors.

Applicable to all of the foregoing is the question: What is the line of demarcation between satisfactory and unsatisfactory performance? Does the standard of satisfactory and unsatisfactory fluctuate as a function of any, or several, of the above?

### Illustrative Criterion Measures

Individual job performance criteria may be classified in many ways, dictated by the institution's goals and organization. For naval officers we have organized them under two broad headings: technical skill and human relations skill.

Each of these sets of skills in turn may be considered as effectors of success at several levels in the operational hierarchy, which in

the Navy would be the ship, the group (departmental, divisional, or other), and the individual.

These may further be sub-classified as to whether they are demonstrated under training conditions or on-the-job.

Finally these criteria may be further specified according to the type of measure being applied, as schematized, for example, in Chapter 5 of R. L. Thorndike's *Personnel Selection*, Wiley, 1949.

---

### How's Your Empathy?

"Empathy" is a word that has been in the dictionaries a long time but it's just beginning to gain recognition as an important quality for executives at all levels. One dictionary definition is "the imaginative projection of one's own consciousness into another being" but as used by psychological consultants in business and industry empathy is used to indicate the ability to imaginatively project the other fellow's consciousness into your own, thereby putting yourself mentally into his shoes, to the point of being able to guess pretty closely what his thoughts and reactions will be in a given situation.

In its simplest form empathy is well illustrated by the old story of the village idiot who found the lost horse when nobody else could. He just sat down and figured out where he would go if he were a horse. He went there and there was the horse.

At a recent meeting of chemical engineers, Dr. Richard S. Schultz, a New York psychological consultant, mentioned the importance of empathy as an executive quality, pointing out that individuals who rate high in this characteristic can more readily understand, predict and control the thinking, feeling, and actions of other people. Psychologists are now at work devising methods of measuring this quality.

"The simplest illustration of empathy is to recall your last experience at an exciting athletic event or theater show," he said. "Remember how you reacted and identified yourself with specific thoughts, feelings, and actions of the feature personalities?"

Empathy, he said, may be further described as a combination of social sensitivity and social intelligence. "It is with such awareness that we can be most skillful in our daily contacts with people," he said.

It is encouraging to know that progress is being made in measuring the characteristic technically known as empathy, for it is a trait that under various inexact tags has been recognized as an important though elusive attribute of success. It is often the reason why two men of apparently equal upbringing, education, intelligence, and opportunity will vary so widely in their degree of business success: one of them can see things from the other fellow's point of view; the other can't, and continually rubs people the wrong way.

If accurate measurements are on the way that will measure this sort of "social savvy" they will be of particular use to the insurance business, in which cooperating with people and getting along with them on a good basis are so much more important than purely technical know-how. (*The National Underwriter*, July 3, 1953.)

## Book Reviews

Maier, Norman R. F. *Principles of human relations, applications to management*. New York: John Wiley & Sons, Inc., 1952. Pp. ix + 474. \$6.00.

This book appears to fulfill three purposes, (1) to present Dr. Maier's research and experience with human relations training programs in industry, (2) to furnish a basic textbook for courses in human relations in industry with material adaptable to laboratory exercises, and (3) to serve as a manual to guide industrial psychologists in introducing human relations programs in business and industry. Although the systematic discussion is based primarily upon Dr. Maier's own work, the conclusions that he reaches are similar to those which have been obtained previously by others in the area of human relations.

The material deals mainly with methods and techniques for a human relations training program, including the use of group discussion methods, role playing, and group decision procedures. The use of such techniques is aimed at overcoming hostility, fears, feelings of insecurity, frustrations, and other barriers to acceptance of democratic supervisory practices. In addition, how to assist the supervisor to be permissive in his dealing with individuals and to use non-directive counseling techniques is discussed at some length. Ample case material is furnished to provide demonstrations of the value of the various methods and techniques discussed. Of particular value is an exposition of how group discussion and role playing techniques can be adapted for use with large groups when it is not possible to use small groups in training.

Concepts employed are well defined and explained. The general ease of reading is marred only by an occasional awkward sentence, and failure to provide adequate transition from one idea to another.

The major emphasis of the book is aimed at explaining and furnishing demonstrations of the value of group discussion and role playing for supervisory training at all levels from top management to line supervisors. Through the use of such techniques it is possible to change a supervisor's feelings and attitudes which conflict with maintaining good human relations with his group of workers. The use

of role playing in training situations directs a supervisor's attention away from the words or logic which are overtly expressed and focuses it upon the feelings which govern the course of interpersonal relations. Most important of course he comes to understand his own involvement in the process. Only through gaining insight in regard to how feelings influence the tenor of interpersonal relations and how they determine the kind of mutual understanding which results can a supervisor, if necessary, come to appreciate and accept new modes of behavior appropriate for dealing effectively with other people. Role playing provides a means of gaining experience under conditions which do not require a supervisor to "save face." Consequently, a situation is provided where he can examine objectively how supervisory attitudes, both good and bad, influence the course and outcome of interpersonal relations. He comes to realize that being permissive is more advantageous for obtaining constructive actions with a concurrent improvement in his status with the workers in regard to their respect for his authority, control and prestige. Such an outcome removes the hampering effect of presumed risks which a supervisor imagines might endanger his capacity to carry out his responsibilities if he adopts democratic practices. Once he is freed of concern for the necessity of protecting his own security, a supervisor is able to let his group solve its own problems under his guidance. When increased effectiveness of the group in accomplishing work ensues, the supervisor is enabled to realize the value of exercising democratic control by utilizing the forces which are in the group rather than by depending upon the use of his power.

Dr. Maier's book is a good illustration of the basic contribution that psychology can make toward developing a realistic philosophy of making life and work tolerable in modern industrial society. True, the basic principles of such a philosophy still go back to Aristotle, Plato, the Christian ethics, and the English common law. But psychology, by utilizing the scientific method, can still contribute materially to their realization by furnishing proof of the effectiveness of various methods and techniques which when fully assimilated

into the mores of our society will come eventually to be considered common sense approaches to maintaining good human relations. It is fortunate that psychology has men with the capability and insight of Dr. Maier, whose approach to developing an applied science of human behavior transcends the bonds of narrow specialization.

Wilton P. Chase

Air Research and Development Command,  
Human Resources Research Center,  
Lowry Air Force Base, Denver, Colorado

Deese, James. *The psychology of learning*. New York: McGraw-Hill, 1952. Pp. x + 398. \$5.00.

Since this book is designed as a text (for advanced undergraduate and graduate students), one of the difficulties inherent in all textbook reviews is encountered here: The professional reader will find in it much that is already familiar, and low in interest value, but which the student may react to very differently. The reviewer must therefore try, as best he can, to look at the book through student eyes. When this is done, the present volume stands up well. It is clearly and simply written, with an occasional colorful turn of phrase. It is, moreover, comprehensive in scope—including, as it does, discussion of animal and human learning in both laboratory and everyday (clinical, applied) settings—and yet wisely avoids trying to be encyclopedic in coverage: "broad rather than exhaustive" is the author's avowed aim. While the book is not founded upon or integrated around any one conception of learning, and thus not so provocative as it might otherwise be, it has the merit of being accurate, critical, and of very possibly inspiring students to get right to work on research designed to fill in the more glaring gaps in our knowledge.

Perhaps the gravest short-coming of the book is the author's failure to interrelate discussions in different chapters. Not infrequently a given piece of research or theory will be discussed, sympathetically and well, in one chapter; and yet, in a later (or earlier) chapter, this material will not be brought to bear upon problems where it is rather obviously relevant. A similar criticism is also occasionally appropriate with respect to ex-

perimental facts and hypotheses available in the literature which have not been included in the book at all. The net result is that the book lacks more in total impact and cogency than it needs to. However, it is a good start and in later editions may well rise to meet the challenge of its topic more fully than it presently does.

The author is well aware of the "applied" potentialities of the psychology of learning and refers from time to time to such fields as education and psychotherapy. However, he is conservative in what he believes laboratory fact and theory can at present contribute along these lines. He very usefully points to some of the not very happy results of premature application of particular conceptions of learning and urges further inquiry rather than rash "practicality."

While *The Psychology of Learning* puts a desirable emphasis upon laboratory procedures as a source of knowledge in this area, it tends to slight what is already known and can be further learned in the "applied" setting. In other words, it does not emphasize as much as it might the reciprocal benefits of interaction between laboratory and field. It sometimes seems to imply that all knowledge originates in the laboratory and is then channeled toward application in the field. The author properly notes that laboratory theories sometimes fall on their face in a practical setting, but he does not, in the reviewer's judgment, give the field proper credit as itself a kind of "laboratory" and certainly a setting which can provide highly stimulating questions and suggestions to be carried back for more rigorous types of investigation.

*The Psychology of Learning* is an excellent job of bookmaking; and despite its being pitched at the textbook level, professional readers will find parts of it novel and exciting.

O. Hobart Mowrer

University of Illinois

Ulrich, David N., Booz, Donald R., and Lawrence, Paul R. *Management behavior and foreman attitude*. Boston: Harvard Business School, 1950. Pp. 56. \$75.

This report is the result of an 8 month case study made by a research team consisting of

the three authors. The study was carried out through informal observation and interviews in a manufacturing firm employing about 500 persons located in a large eastern city. About half the time of the study was spent in observation of an assembly department of 36 female employees and their foreman.

As the title implies, the main object of the study was to determine what effects the behavior of top management had on the foreman. In addition, the effects of the behavior of other groups on the foreman, including his employees, staff specialists, and his immediate superior were also studied.

A number of difficult and strained relationships at all levels of the organization are pointed out, causes of these difficulties hypothesized, and recommendations made on how the relationships could be improved. A major recommendation made is that top management make greater efforts to understand the effects of administrative action on employees and supervisors.

Because the only evidence given to back up what is said consists of scattered selected quotations of remarks made by those observed, the reader will find he is being asked to accept the conclusions and recommendations pretty much on faith in the analytical ability of the researchers. Despite this limitation, however, few readers who deal with similar problems will finish this report without some new insights into these problems and new ideas for dealing with them in their own situations.

Theodore R. Lindbom

*Personnel Department,  
Midland Cooperative Wholesale,  
Minneapolis, Minn.*

Weinland, James D. and Goss, Margaret V. *Personnel interviewing*. New York: Ronald Press, 1952. Pp. vii + 416. \$6.00.

This book deals with the aims and techniques of business interviewing and is ad-

dressed to individuals concerned with personnel relations and employment. Although many of the principles and procedures are applicable to all types of personnel interviewing, the book emphasizes employment. Chapters are devoted, however, to other types of interviewing, such as merit rating, disciplinary, counseling, etc.

A section on the interviewer and his work deals with introductory and background material, ranging from individual differences to interviewing environment and the training of interviewers. A second part deals with techniques, including material on directive, non-directive, and patterned interviews. The third part of the book deals with interviews for various purposes.

Although the book contains much of value and has interesting material and views, the over-all effect is disappointing. Perhaps one reason for disappointment is the great need for a comprehensive and up-to-date text in the field of personnel interviewing. The chapters of this book get off to a good start, but the reviewer had a feeling of disappointment at the end of each. This was due partly to failure of the authors to organize and systematize the material adequately. This is true in spite of their predilection for lists and classifications. Unfortunately, such lists often appeared incomplete or haphazard.

The authors are guilty of looseness, ambiguity and over-generalization. One suspects that some dogmatically worded statements would be considered better as hypotheses than as proved facts.

The reviewer's over-all opinion is indicated by the fact that although he is currently training interviewers, he is not using this book. Other materials are being used even though older, or available only in less accessible form.

Clifford E. Jurgensen

*Minneapolis Gas Company*

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- The workshop handbook.* Walter A. Anderson, Rollin P. Baldwin, and Mary Beauchamp. New York: Columbia University, 1953. Pp. 65. \$1.00.
- Adjustment to physical handicap and illness: A survey of social psychology of physique and disability.* Roger G. Barker, Beatrice A. Wright, Lee Meyerson, and Mollie R. Gonick. New York: Social Science Research Council, 1953. Pp. 440. \$2.00.
- Differential migration in the corn and cotton belts.* Donald J. Bogue and Margaret Jarman Hagood. Oxford: Scripps Foundation, 1953. Pp. 248. \$2.25.
- The fourth mental measurements yearbook.* Oscar K. Buros, Editor. Highland Park, N. J.: The Gryphon Press, 1953. Pp. 1,189. \$18.00.
- Group dynamics.* Dorwin Cartwright and Alvin Zander. Evanston: Row, Peterson and Company, 1953. Pp. 642.
- Human behavior: psychology as a bio-social science.* Lawrence E. Cole. New York: World Book Company, 1953. Pp. 884. \$4.56.
- A factor analysis of verbal and non-verbal tests of intelligence.* Reverend James T. Curtin. Washington, D. C.: The Catholic University of America Press, 1952. Pp. 63. \$1.25.
- Raising the sights of office management.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 59. \$1.25.
- Industry enters the atomic age.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 31. \$1.25.
- Guides to meeting tomorrow's production needs.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 64. \$1.25.
- Planning for worker security and stability.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 40. \$1.25.
- The new climate of union-management relations.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 32. \$1.25.
- Factors in intelligence and achievement.* Justin A. Driscoll. Washington, D. C.: The Catholic University of America Press, 1952. Pp. 56. \$1.00.
- College board scores.* Henry S. Dyer. New Jersey: College Entrance Examination Board, 1953. \$.75.
- Stabilization of employment is good management.* Charles C. Gibbons. Kalamazoo: W. E. Upjohn Institute for Community Research, 1953. Pp. 16. Gratis.
- The uneducated.* Eli Ginzberg and Douglas W. Bray. New York: Columbia University Press, 1953. Pp. 246. \$4.50.
- Psychosis and civilization.* Herbert Goldhamer and Andrew Marshall. Glencoe: The Free Press, 1953. Pp. 126. \$4.00.
- Measurements of human behavior.* Edward B. Greene. Revised edition. New York: The Odyssey Press, Inc., 1953. Pp. 790. \$4.75.
- A clinical approach to children's Rorschachs.* Florence Halpern. New York: Grune and Stratton, Inc., 1953. Pp. 288. \$6.00.
- Introduction to psychology.* Ernest R. Hilgard. New York: Harcourt, Brace and Company, 1953. Pp. 659. \$7.50.
- Current problems in psychiatric diagnosis.* Paul H. Hoch and Joseph Zubin. New York: Grune & Stratton, 1953. Pp. 291. \$5.50.
- The psychology of successful selling.* Richard W. Husband. New York: Harper and Brothers, 1953. Pp. 306. \$3.95.
- Techniques of successful foremanship.* Eugene E. Jennings. Madison: University of Wisconsin, School of Commerce, Bureau of Business Research and Service, 1953. Pp. 41. \$1.15.
- Psychology and alchemy.* C. G. Jung. New York: Bollingen Foundation, Inc., 1953. Pp. 563. \$5.00.

- The psychology and psychotherapy of Otto Rank.* Fay B. Karpf. New York: Philosophical Library, 1953. Pp. 129. \$3.00.
- Elementary school objectives.* Nolan C. Kearney. New York: Russell Sage Foundation, 1953. Pp. 189. \$3.00.
- Rehabilitation of the physically handicapped.* Henry H. Kessler. New York: Columbia University Press, 1953. Pp. 275. \$4.00.
- Statistical methods in experimentation.* Nolan C. Lacey. New York: Macmillan Co., 1953. P. 249.
- The retarded reader in the junior high school.* May Lazar, Editor. New York: Board of Education, Bureau of Educational Research, 1952. Pp. 126.
- The psychology of personal and social adjustment.* Henry Clay Lindgren. New York: American Book Company, 1953. Pp. 481. \$4.50.
- Design and analysis of experiments in psychology and education.* E. F. Lindquist. Boston: Houghton Mifflin Co., 1953. Pp. 393. \$6.50.
- In the minds of men.* Gardner Murphy. New York: Basic Books, Inc., 1953. \$4.50.
- Rorschach interpretation: advanced technique.* Leslie Phillips and Joseph G. Smith. New York: Grune and Stratton, Inc., 1953. Pp. 400. \$8.75.
- Wait the withering rain.* Austin L. Porterfield. Fort Worth: Leo Potishman Foundation, 1953. Pp. 147. \$2.50.
- Social psychology.* S. Stansfeld Sargent. New York: The Ronald Press Company, 1953. Pp. 519. \$4.50.
- Occupational information.* Carroll L. Shartle. New York: Prentice-Hall, Inc., 1952. \$5.00.
- An experiment in recreation with the mentally retarded.* Bertha E. Schlotter and Margaret Svendsen. Chicago: Illinois Department of Public Welfare, 1951. Pp. 142. Gratis.
- Medical public relations.* Edgar A. Schuler, Robert J. Mowitz, and Albert J. Mayer. New York: Health Information Foundation, 1952. Pp. 228.
- Groups in harmony and tension.* Muzafer Sherif and Carolyn W. Sherif. New York: Harper & Brothers, 1953. Pp. 316. \$3.50.
- Introduction to experimental method.* John C. Townsend. New York: McGraw-Hill Book Co., Inc., 1953. \$4.00.
- Modern educational problems.* Arthur E. Traxler, Editor. Washington, D. C.: American Council on Education 1953. Pp. 147. \$1.50.
- Improving transition from school to college.* Arthur E. Traxler and Agatha Townsend. New York: Harper & Brothers, 1953. Pp. 165. \$2.75.
- Personality tests and assessments.* Philip E. Vernon. London: Methuen & Co. Ltd., 1953. Pp. 220.
- The roots of psychotherapy.* Carl A. Whitaker and Thomas P. Malone. New York: The Blakiston Company, Inc., 1953. Pp. 236. \$4.50.
- The measured effectiveness of employee publications.* Association of National Advertisers, 1953. Pp. 109.
- Drug addiction among adolescents.* Committee on Public Health Relations of the New York Academy of Medicine. New York: The Blakiston Company, Inc., 1953. Pp. 320. \$4.00.

# Journal of Applied Psychology

VOL. 37, No. 6

DECEMBER, 1953

## The Prediction of Proficiency of Taxicab Drivers

Clarence W. Brown and Edwin E. Ghiselli

*University of California, Berkeley*

In the evaluation of devices for use in the selection of operators of public conveyances, greatest attention has been given to the criterion of accidents. In some instances labor turnover has been considered, but safety of performance has been given greater emphasis. While the importance of accidents and labor turnover certainly is not to be minimized, it should be apparent that the success of operators of vehicles can be measured in other important ways. In the taxicab industry, for example, job success can be gauged in terms of the dollar volume of business that the driver achieves. The economic health of a taxicab company can be improved by reducing costs due to accidents and personnel replacements, but it is more directly related to the monetary return accomplished from the sale of its services. It is apparent, therefore, that the selection of individuals who can sell their services as taxicab drivers is worthy of consideration.

Very little information is available on the effectiveness of predicting the productivity of taxicab drivers. Wechsler reports inconsistent correlations between sales and intelligence test scores (4). Viteles found intelligence tests to be of no value but obtained substantial predictions from a weighted personal data blank (3). The results of investigations conducted to date provide little help in planning an experimental program for driver selection.

### Criteria

The amount of business conducted by a taxicab company is subject to a number of uncontrolled variables. It is affected by such obvious factors as weather and season. But in addition, sales are sensitive to other types

of occurrences such as large civic entertainments, conventions, the payment of bonuses by some large local organization, and the like. In many instances sales rise or fall for no discernible reason. Since these variations may be as great as 100%, it is apparent that corrections must be applied to a driver's sales in order to compensate for the time trends in the volume of business. In the present investigation the average sales for all drivers were computed for each week, and the productivity of each driver was expressed as a percentage of this average. This procedure controlled most but not all of the time trends. These weekly indices formed the basis of the production criteria employed in the validation studies reported here.

One characteristic of the taxicab industry which is pertinent to all selection studies is the high rate of turnover among the drivers. A person who has been with a company for a year is considered to be an "old hand." This high labor turnover means that production records for any extensive periods of employment are not obtainable for large numbers of drivers working under relatively homogeneous conditions. In the present investigation production during the first eighteen weeks of employment was used. In spite of the fact that this period of time is relatively short, the reliability of the measures of proficiency was quite satisfactory. The coefficient of correlation between production indices on odd and even weeks, corrected by the Spearman-Brown formula, was found to be .96.

A cross validation study of the tests was conducted in a second and smaller company. Due to the particular accounting methods of this company, sales records were not avail-

able in a usable form. The manager of the company, however, provided ratings of his drivers' productivity on a six-point scale. In making the ratings the manager discussed each driver with the investigators, thus carefully reviewing the driver's achievement before placing him in one of the rating categories. While no evidence of reliability was obtained, in terms of distribution statistics, at least, the ratings were satisfactory. The ratings were made on the men after three months of employment.

### Subjects

The subjects in the present investigation were men who applied for work as taxicab drivers and who were hired. Only those cases were used who had no previous experience in driving taxicabs. They did vary, however, in the amount of experience they had had in driving other types of commercial vehicles.

Various selective factors operated so that the subjects used were by no means representative of the entire range of talent of applicants. Prior to being hired the men were interviewed and took a driver road test. About 20% were rejected on these bases. In addition, about another 20% were rejected on the basis of very poor scores on the aptitude tests to be described here. As a final selective factor, only those cases were used who remained on the job either 18 or 12 weeks or more. In the two companies used in the present investigation, approximately 40% of the drivers left their jobs within the 18 or 12 week periods. The men utilized in this study, then, represent about 20% of the applicants; those who survived the hiring procedures and remained on the job at least 18 weeks for the larger company and 12 weeks for the smaller company. For the basic validation study, 54 men were drawn from the first company, and for the cross validation study 29 men were drawn from the second company.

### Predictor Variables

Seven aptitude tests were utilized together with an interest inventory. All of the tests were time limited, and all were of the paper and pencil variety. As indicated earlier, the tests and the inventory were administered

prior to hiring. The choice of the particular measures utilized was dictated by an interest in predicting several aspects of success rather than concentrating on sales alone. Certain of the measures were found to predict accidents and labor turnover (1, 2).

An arithmetic test was employed which involved problems in making change and computing fares. A test, termed Speed of Reactions, presented the individual with a series of rules that he was to use in making differential responses to various spatial arrangements and organizations of letters. Some indication of motor speed and precision was obtained from dotting and tapping tests. The dotting test called for the placing of a single dot in each of a series of irregularly spaced circles. In the tapping test only speed was required, the individual tapping as rapidly as possible with his pencil, placing three dots in each of a series of circles.

Two tests of spatial ability were administered which primarily involved the ability to detect differences in distances. In the Judgment of Distance test each item was a schematized table top on which rested four cubes of equal size. On the basis of perspective and interposition the individual judged which cubes were nearest together. The Distance Discrimination test called for the discrimination of linear distances between points. A Mechanical Principles test was used which consisted of a series of pictorially presented problems each of which required knowledge of some simple principle of mechanics.

In the interest inventory each item involved a pair of occupations or jobs, and the individual chose the one of each pair which he preferred. The choices were between a higher and a lower occupation, a job performed outside as compared with one performed inside, a job involving dealing with people rather than one not requiring such activity, and a job involving moving about rather than one requiring sedentary activity.

### Results

Table 1 gives the validity coefficients for the various predictor variables using the sales production criterion for the basic group of 54 drivers. With the possible exception of

Table 1

Validity Coefficients of Several Tests for Predicting Sales Production of 54 Taxicab Drivers

Test	Validity Coefficient
Arithmetic	.29
Speed of Reaction	-.19
Dotting	.21
Tapping	.18
Judgment of Distance	-.03
Distance Discrimination	.24
Mechanical Principles	.13
Interest Inventory	.20

the arithmetic test, none of the predictors alone would be considered to give adequate prediction. When the extent of restriction in range of talent is considered, however, low coefficients assume some importance. With the exception of the Judgment of Distance test, all measures would seem to merit further study.

A simple combination of test scores was effected by eliminating the Judgment of Distance test, assigning unit weight to each of the others, and assigning a negative value to the Speed of Reaction test. In effect, this composite score was the sum of the standard scores of the individual tests. The validity of this battery score for the 54 basic cases was .39, which is a reasonably satisfactory prediction.

As is well known, there is almost always a shrinkage in validity coefficients in cross validation studies. The test weights mentioned above were used in validating the scores of the 29 cases in the second company. In this cross

validation the validity of the battery was found to be .29. While this value may not appear to be particularly significant it is to be remembered that in addition to restriction of range of talent this coefficient is affected by the use of a somewhat different criterion.

In view of the complex nature of the production criterion, it is surprising that such tests as dotting, tapping, and discrimination of distances have any predictive power at all. No logic would lead an investigator to employ such tests in predicting sales of taxicab service. To be sure, the extent of prediction by individual tests was low, but the combination of tests gave a usable index of aptitude.

### Summary

Seven tests and an interest inventory were administered to 54 taxicab drivers and validated against their sales. With one possible exception, no single test gave adequate prediction. A simple weighted combination of the tests yielded a validity of .39. When the weighted battery was applied to another group of 29 drivers it was found to have a validity of .29 in the prediction of ratings of job proficiency.

Received March 16, 1953.

### References

1. Brown, C. W. and Ghiselli, E. E. Prediction of labor turnover by aptitude tests. In press.
2. Ghiselli, E. E. and Brown, C. W. Prediction of accidents of taxicab drivers. *J. appl. Psychol.*, 1949, 33, 540-546.
3. Viteles, M. S. *Industrial psychology*. New York: Norton, 1932.
4. Wechsler, D. Tests for taxicab drivers. *J. Person. Res.*, 1926, 5, 24-30.

## Some Measured Characteristics of Air Force Weather Forecasters and Success in Forecasting<sup>1</sup>

James J. Jenkins

*University of Minnesota*

A review of the psychological and meteorological literature reveals that little is known about the measured psychological characteristics of weather forecasters, and this writer has found no studies relating such characteristics to occupational success. In the past it appears that high scholastic ability or achievement has been accepted as essential. Selection practice in the AAF Technical Training Command during World War II stressed high scores on tests of academic ability, mathematics, and physics (e.g. 10, 11, 12) since the Weather Forecasting course was believed to be one of the most difficult courses offered in the technical training schools. Success in the course showed low positive correlations with the AGCT and mathematics tests (e.g. 8, 9). Harrell (2) in a survey of AGCT scores of 209 AAF technical specialties found the enlisted weather forecasters to be the highest ranking group with a median score of 136.7. This perhaps indicates only that the screening on intelligence was very effective.

The purpose of the present study was: (1) to determine how Air Force forecasters are differentiated from a more general population; and (2) to disclose the extent to which certain measures are associated with ability to forecast weather.

### Procedure

In 1948 the writer secured the cooperation of the Air Weather Service for a study of some of the psychological characteristics of forecasters and the possible relation of these

<sup>1</sup> This study was made possible by the cooperation of the Air Weather Service, U. S. Air Force. It was undertaken with the encouragement of General D. N. Yates, then Chief of the Air Weather Service. The writer is especially indebted to Prof. Donald G. Paterson for his assistance and guidance in every phase of the study. This paper is part of the writer's Ph.D. thesis on file in the library of the University of Minnesota under the title of "Prediction of forecasting efficiency for Army weather forecasters." 1950.

characteristics to success in forecasting. A study of available job descriptions and lists of qualifications (e.g. 13, 14) and a job analysis from these sources and the writer's own experience as a forecaster resulted in the selection of the following variables for consideration as related to success: education, college major, mathematics background, forecasting and observing experience, kind of meteorological training, forecasting aids most frequently used, speed and accuracy of perception, spatial relations ability, general academic ability, and vocational interests. Information on all but the last four of these was gathered by means of a questionnaire. The remaining variables were measured by the Minnesota Clerical Test, the Revised Minnesota Paper Form Board, the Ohio State University Psychological Test, and the Strong Vocational Interest Blank for Men. The tests were administered to the forecasters by the Air Weather Service and the results returned to the writer.

### The Criterion

The problem of obtaining criterion data had been encountered by the Air Weather Service early in World War II. Muller (5) in a review of the literature on verification of forecasts points out that no less than 54 methods of evaluation were proposed between 1893 and 1943 and that all of these have been vigorously criticised. After a long program of experimentation by the Weather Information Branch, a special verification method was devised by Lt. M. J. Slonim (15) which seemed to avoid most of the usual difficulties. This procedure consisted of evaluating the probabilities of occurrence of given values for each forecast element (pressure, temperature, precipitation, visibility, and ceiling) from climatological data for the time and location being forecast. A scale of 30 equal probability units (or *trentiles*) was set up by which observed and forecast values could be

compared. Discrepancies were summed in probability units to indicate the relative accuracy of forecasts. For example, to score pressure forecasts for a given station at a given time, we would proceed as follows: (1) gather past observations for this period of the year for this locality and make a frequency distribution of observed pressures; (2) divide this frequency distribution into 30 intervals of equal or nearly equal probability of occurrence (see Table 1 for a partial example); and (3) score forecasts now obtained in terms of the number of equal probability intervals (or trentiles) in the discrepancy between the forecast and observed values. (In our example, if a pressure of 995 is observed and a pressure of 1000.8 is forecast, the score is zero. If the forecast is 1003.0 the score is one. If the forecast is 1033.0 the score is thirty, etc.)

Each Air Force forecaster in the United States was required to make at least three forecasts per week for five widely scattered

stations selected by the Weather Information Branch. All forecasts were made from the 1230 Greenwich time maps. The data available to the forecasters were approximately the same regardless of their location in the country.

This program ran from 1943 to 1945 and furnished the criterion data for this "post-diction" study. The criterion yielded a reliability of .90 (estimated from a part-whole correlation of .70 between 8 weeks of the program and the total 84-week program). It is unfortunate that these data are now available only in terms of standard scores so that the relative accuracy of the forecasts in terms of initial values is unknown. The homogeneity or heterogeneity of the forecasters as a group is impossible to assay even in terms of probability deviations. It is also regrettable that the conditions under which the forecasts were made (time pressure, amount of other work, freedom from interference, etc.) could not possibly be equated. The validity evidence is largely "face validity" (15).

Table 1

A Hypothetical Frequency Distribution of Barometric Pressures and the Resulting Trentile Table for Station "X" for a Given Thirty-Day Period

Pressure Distribution		Trentile Table	
Pressure in Millibars	Number of Observations	Values of Element	Trentiles
998.7	1	Less than and including 1001.2	1
999.4	1		
1000.3	1		
1000.8	1		
1001.2	1		
1001.7	1	From 1001.3 to 1003.9 inclusive	2
1003.1	1		
1003.7	1		
1003.9	3		
etc.			
1032.2	2	From 1031.8 to any greater value	30
1033.5	1		
1033.9	2		

Note: The middle portion of this table is omitted because of excessive length.

## The Sample

The sample for this study was sharply restricted by three conditions. First, the forecaster must have participated in the criterion study. Second, he must have remained in the Air Weather Service until 1948. Third, in 1948 he must have been stationed in or near the United States so that he could be tested. Only 92 forecasters met all these conditions and constituted the sample for this study. The forecasting scores of the sample were compared to those of the total group participating in the verification program ( $N = 2023$ ) and were found to resemble them closely ( $\chi^2 = 2,832$ ;  $P = .88$ ).

## Characteristics of the Sample

*General.* All but two of the forecasters graduated from high school, but only 30 had graduated from college. Three had Ph.D. degrees and 12 others had done post-graduate work. Average education was 14.3 years with a standard deviation of 2.2 years. A total of 57 indicated college majors and of these 49 were in the natural sciences or mathematics. Before starting meteorology training the aver-

age number of college mathematics courses was three. The range was from no courses to two Ph.D.'s in mathematics. All of the sample had received training in meteorology in military schools under contract with the Air Forces. Of the sample, 71 per cent had previous experience as weather observers.

*Test Data.* Means and standard deviations in raw scores for the ability tests are given in Table 2. Reference to relevant norm groups reveals the forecasters to be a highly selected group on all of these variables.

Table 2

Means and Standard Deviations of Forecasters  
on Ability Tests

Tests	Mean	Standard Deviation
Minnesota Clerical Test		
Numbers	141.1	29.8
Names	145.8	31.9
Revised Minnesota Paper Form Board	50.1	7.5
Ohio State University Psychological Exam.		
Part I	24.5	3.4
Part II	44.6	10.2
Part III	47.2	6.8
Total	116.3	18.1

The mean scores on the Minnesota Clerical Test fall at the 95th and 93rd percentiles for Numbers and Names sections respectively when compared to gainfully occupied adults and at the 60th and 73rd percentiles when compared to employed clerical workers themselves (1).

The mean score on the revised Minnesota Paper Form Board when compared to the norms of various male industrial groups (3) falls from the 80th to the 97th percentile with a median value at the 90th percentile. Even compared to first and fifth year engineering students the percentile ranks are 80 and 70 respectively.

For the Ohio State University Psychological Test the forecasters were compared with college freshmen as a group which roughly approximated the pre-army educational status of the sample. On this basis the mean for

Table 3

Means, Standard Deviations, and Percentage of A and B+ Ratings for Each Strong Key for Total Sample of 92 Weather Forecasters

Group	Occupation	Mean	Standard Deviation	Per Cent A & B+
I	Artist	19.8	10.0	5
	Psychologist	19.4	12.4	12
	Architect	28.3	10.5	11
	Physician	28.7	9.7	11
	Osteopath	32.2	9.7	28
	Dentist	27.9	9.9	14
II	Mathematician	25.5	9.9	10
	Physicist	22.7	13.8	11
	Engineer	40.8	10.1	58
	Chemist	38.2	11.5	49
III	Production Manager	42.6	8.2	65
IV	Farmer	40.3	9.2	60
	Aviator	43.2	10.2	71
	Carpenter	31.7	10.9	24
	Printer	39.7	8.9	54
	Math. Phys. Sci. T.	43.3	9.7	67
	Policeman	36.9	8.7	40
	Forest Service Man	35.1	9.5	33
V	YMCA Phys. Director	30.5	9.6	21
	Personnel Director	37.0	10.7	42
	Public Admin.	43.3	8.7	66
	YMCA Secretary	22.7	10.1	4
	Soc. Sci. H. S. T.	31.1	10.5	26
	City School Supt.	23.4	9.1	2
	Minister	18.5	11.1	4
VI	Musician	28.5	10.9	20
VII	C. P. A.	26.0	8.8	7
VIII	Accountant	36.0	10.6	35
	Office Man	37.8	9.6	46
	Purchasing Agent	34.2	8.9	24
	Banker	27.5	8.6	5
	Mortician	27.0	8.8	9
IX	Sales Manager	28.1	9.4	11
	Real Est. Sales.	31.2	6.6	8
	Life Insur. Sales.	22.1	10.5	5
X	Advertising Man	26.8	7.9	5
	Lawyer	25.8	8.1	3
	Author-Journalist	25.6	7.5	4
XI	Pres.-Mfg. Concern	29.8	7.5	11
	Interest Maturity	55.2	5.5	—
	Occupational Level	52.6	6.4	—
	Masc.-Fem.	54.6	8.4	—

the forecasters falls at the 87th percentile in Part I (Same-opposites), 81st in Part II (Word relationships), 90th in Part III (Reading comprehension), and 87th for the total score (7).

It is readily apparent that the forecasters are a superior group on all three of these tests. While this superiority might be expected on the Ohio in view of the initial screening, their superiority on the other two tests is not readily explained.

The means and standard deviations for each of the Strong Vocational Interest Blank keys are given in Table 3 in terms of the occupational standard scores. The high mean scores of the meteorologists (B+ and B) show their interests to be similar to those of persons in the occupations of Engineer, Chemist, Production Manager, Farmer, Aviator, Printer, Mathematics-Physical Science Teacher, Policeman, Forest Service Man, Personnel Director, Public Administrator, Accountant, Office Man, and Purchasing Agent. Their interests are most markedly *dissimilar* (scores in the C area) to those persons in the occupations of Actor, Banker, Mortician, Real Estate Salesman, Life Insurance Salesman, Advertising Man, Lawyer and Author-Journalist. Most of the rest of the scores were in or near the chance range.

If one views these occupations in terms of Strong's factor analysis data (6), the similarity of this grouping of the occupations to Factor III ("Language" or "things *versus* people") is immediately apparent. All of the occupations in which the meteorologists score high have positive loadings on this factor (in the direction of "non-language" and "things") and all the occupations in which they score low have negative loadings (in the direction of "language" and "people").

The picture of the forecasters seen in the interest test results is one of a technical, skilled-trades interest group with little verbal-linguistic, pure science, or social service interest. The relatively low OL score received by the group seems to reflect the technical skilled-trades kind of picture already given.

### The Prediction of Forecasting Ability

In order that the findings might be subjected to cross-validation the sample was split in half. Individuals were paired on criterion scores and for each pair a random determination was made as to which member fell in the first group and which into the second group. A double cross-validation technique (4) was used, prediction devices being prepared on each group and validated on the other group. Regression, cutting scores, and profile techniques were utilized. The final results of this procedure are summarized here rather than presenting the work in detail.

No consistent differences were found between the better and poorer forecasters with respect to age, rank, education, college major, mathematics background, forecasting and observing experience, kind of training, forecasting aids most frequently used, interest test profiles or scores on the Revised Minnesota Paper Form Board. The Ohio State University Psychological Test proved to be of little use in discriminating degrees of success in forecasting but of the 27 persons who scored low (below 44) on Part III (reading comprehension), 19 of them fell in the lower half of the forecasting group. High scorers were not, however, distinguished from other forecasters.

The Numbers section of the Minnesota Clerical Test proved to be of no value in the prediction of forecasting accuracy but the Names section proved to be of considerable value. In both halves of the sample it correlated +.31 with the criterion. When used alone with a cutting score it consistently eliminated at least twice as many cases from the lower half of the criterion group as from the upper half. When used in profile relationship with the Numbers section of the test and the Revised Minnesota Paper Form Board, it eliminated 35 per cent of the cases in the lower half and only 4 per cent of the cases in the upper half. (This amounts to a crude use of the other two tests as suppressor variables. They correlate positively with the Names section and essentially zero with the criterion.)

## Discussion

In view of the findings of this research it would seem that the role of speed and accuracy in perception as measured by some component of the Names section of the Minnesota Clerical Test should be investigated carefully in future studies of weather forecaster training and success on the job. The high level of clerical ability found in the forecasters as a group seems to argue that some kind of selection on this variable is already taking place, and the correlation with forecast verification seems to indicate that this is an important though not an *a priori* obvious source of variation in on-the-job performance.

It should be noted, however, that the search for predictors cannot be considered at all complete. Both of the tests which proved to have any predictive efficiency in this study functioned only at the lower score levels to provide negative selection. A study of other abilities and the motivational and personality characteristics of those individuals who were high on all of the tests employed here but still relatively low in forecasting accuracy is obviously necessary.

## Summary

A sample of 92 Air Force Weather Forecasters was studied to determine: (1) how the sample was differentiated from a more general population; and (2) the extent to which biographical data and psychological measures were associated with ability to forecast. Subjects completed a questionnaire and four standard psychological tests. The sample proved to be similar to the World War II population of forecasters with respect to forecasting ability as measured by the Short-Range Forecast Verification Program.

The sample proved to be a highly select group with respect to educational background, clerical ability, spatial relations ability, and general academic ability. With respect to interests the forecasters appear to resemble a technical, skilled-trades interest group with little verbal-linguistic, pure science, or social service interests.

A double cross-validation study to predict forecasting accuracy revealed only one consistent predictor, the Names section of the Minnesota Clerical Test, which correlated + .31 with skill in forecasting.

It is suggested that further studies of the role of perceptual skills and personality variables in weather forecasting are needed.

Received February 6, 1953.

## References

1. Andrew, Dorothy M. and Paterson, D. G. *Minnesota Clerical Test Manual*. New York: Psychological Corporation, 1946.
2. Harrell, T. W. Army General Classification Test results for Air Forces specialists. *Educ. psychol. Measmt.*, 1946, 6, 341-349.
3. Likert, R. and Quasha, W. H. *The Revised Minnesota Paper Form Board Test Manual*. New York: Psychological Corporation, 1948.
4. Mosier, C. I. Problems and designs of cross-validation. *Educ. psychol. Measmt.*, 1951, 11, 5-11.
5. Muller, R. H. Verifications of short-range weather forecasts—a survey of the literature. I, II and III. *Bull. Amer. Meteorological Soc.*, 1944, 25, 18-27; 47-53; 88-95.
6. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford University Press, 1943.
7. Toops, H. A. *Manual of Directions; The Ohio State University Psychological Test*. Chicago: Science Research Associates, 1941.
8. *Air Forces Technical Schools Validation Study*. AAF Technical Training Command. 9-19-42.
9. *Air Forces Technical Schools Validation Study*. AAF Technical Training Command. 9-30-42.
10. *Classification Division Bulletin*. No. 7. Knollwood Field, N. C., AAF, Hq., Technical Training Command. June, 1942.
11. *Classification Division Bulletin*. No. 13. Knollwood Field, N. C., AAF, Hq., Technical Training Command. Sept., 1942.
12. *Classification Division Bulletin*. No. 16. Knollwood Field, N. C., AAF, Hq., Technical Training Command. Dec., 1942.
13. *Meteorology as a Profession*. Vocational Booklet No. 4. National Roster of Scientific and Specialized Personnel. Washington: United States Government Printing Office, 1947.
14. *Physical Sciences*. Description of Professions Series Pamphlet No. 6. National Roster of Scientific and Specialized Personnel. Washington: United States Government Printing Office, 1947.
15. *Short-Range Forecast Verification Program*. Technical Report 105-26. Publications of the Weather Information Branch. Hq., AAF, 1943.

## A Note on Small Samples

Edward N. Hay

*Edward N. Hay & Associates, Inc., Philadelphia, Pa.*

The sample presents more delicate problems than almost any other aspect of measurement. To begin with, many psychologists working on problems of testing, have been in the habit of going through many refined operations to correct for the deficiencies of a "small sample." Small sample methods were developed in agronomy, where it is possible to hold most of the variables reasonably constant. This is much less true in testing human beings. Consequently, the mere application of small sample statistics cannot be expected to produce automatically more valid results than would be the case without them. Sometimes the characteristics of a sample are such that no amount of treatment will bring about a satisfactory result.

The ceaseless search for "large samples" has resulted in many errors. A psychologist not long ago, in the course of an industry study, published norms which were the result of adding together a great many small samples. It was not possible to make any check on the soundness of this operation because of lack of information. However, large samples have frequently been made out of groups of small samples, when the characteristics of individual small samples differed widely. In one instance, the mean of one sample was more than one sigma away from the mean of another sample. The reasons for the difference could only be conjectured but certainly one larger sample was not to be

had by adding two small samples which differed this much.

Some years ago, I violated my own principles by assembling a group of 120 subjects from more than 20 departments of a single company. The resulting "hash" actually produced reasonably satisfactory validity coefficients just by luck. On another occasion, there were three departments of 10, 7, and 24 employees respectively. Before adding them together to make a "large" sample, an examination of the characteristics of all three groups was made. This revealed that there was little variability in the test scores for the two smaller groups. In the circumstances, neither one could produce a validity coefficient or contribute to one when added to other samples. The larger group gave  $r$ 's exceeding .5 on a number of tests.

Bransford has developed procedures for handling the criterion measures so as to be able to combine samples which are unlike in some degree; for example, ratings made by different raters. He spoke on this topic before the Eastern Psychological Association at Atlantic City in 1952 under the descriptive title "Summational Within-Group Analysis." Combining criterion measures presents more difficulties usually than combining the scores of tests of the same groups.

*Received September 17, 1953.*

*Early publication.*

## The Measurement of Personality and Behavior Disorders by the I. P. A. T. Music Preference Test

Raymond B. Cattell and Jean C. Anderson

*Laboratory of Personality Assessment and Group Behavior, University of Illinois*

On the wide research front which is roughly designated by "projective tests," but perhaps more accurately by "misperception tests" (1), few recent advances have been so promising as that connected with music perception. The powerful and immediate connection of musical stimulation with emotional experience, and the many indications that unconscious needs gain satisfaction through this medium, have long pointed to measures of musical preference as effective avenues to deeper aspects of personality. Moreover, the lack of verbal content is itself, on general principles, a promise that the verbal, cognitive defenses of the censor may be by-passed and the emotional needs probed more directly without distortion by defense elaborations.

### The Music Preference Test

Personality tests which proceed from the esthetic reactions of the subject, or from likings and dislikings which cannot be based on logical, explicit relationships to the subject's purposes and sentiments, occupy an area intermediate between that of projective tests and that of other objective personality tests. For the liking or disliking is evidently due to characteristics imported or projected into the physical sounds by the listener, yet the "projections" are not so explicit as in the imagery evoked by the Rorschach or the interpretive stories which the subject is asked to weave around the T.A.T. It is possible, therefore, that further research and clinical experience with this relatively unexplored class of tests (which may be called tests of "affective misperception") will show them to have certain advantages over the standard projective or misperception tests. For sophisticated subjects intuitively realize that their *cognitive* projections stand in need of defensive disguise, whereas their likings and dislikings make no more sense to them than they do to

the psychologist—before his statistical analyses are made.

As in all test construction involving "items," it would be foolish here to design psychological measures hinging on the luck of a single response and to attempt to relate such a single response to personality dimensions. Instead we first seek reliability in the test measurement itself by composing it of scores on several items, thereby diminishing the effects of chance and specific historical associations. This may be conceived as discovering the dozen or more "items" that can *be validly added together to give a score on some single dimension of emotional quality or musical-emotional reactivity*. Attempts to find these groupings by introspection or by psychiatric judgments must be set aside, for they are shown by preliminary research to be highly unreliable and to constitute an amateurish approach to the problem. Instead it is necessary to find the dimensions of musical choice by submitting a number of musical excerpts to a large population and correlating the responses, thereby discovering empirically which responses "go together." This first stage of research in the area has already been carried out by Cattell and Saunders (4) using 120 half-minute musical excerpts under conditions described elsewhere.

The psychologically interesting and reassuring thing about this factor analysis of a matrix couched in a new variety of response correlations, namely, in music preference responses, is that simple structure was as definitely obtained here as with ability tests, and that a comparison of two factorizations revealed a very gratifying degree of invariance of the factors. With this assurance from an initial study it is to be hoped that psychologists will be encouraged to face the vast amount of exacting work required by this approach instead of being beguiled by merely esthetic intuitions in test construction.

The two dovetailed factor analyses yielded eleven stable factors (4). But before these basic findings could become a practical foundation upon which further "applied" research could readily go forward, in clinics and guidance centers generally, it was first necessary to construct out of the above research findings a convenient routine instrument. This was done under the auspices of the Institute for Personality and Ability Testing by the senior author and has issued in a 12-inch long-playing record, reproducing 100 half-minute music excerpts (50 on one side, Form A, and an equivalent 50 on the other, Form B). Except for the first and the last three factors in this test there are ten items provided to measure each factor. These items were chosen from the 120 factorized, according to the usual test construction principles: a significant loading on the factor concerned; a balancing (suppression) of loadings on factors not concerned; a balancing of "like" and "dislike" responses in the score for any one factor; no use of any item for more than one factor. A cyclical order of sampling of items from the various factors is used in the test as finally presented.

The test so constructed, when cross validated on a new population, was found to have consistency (split-half reliability) and equivalence (Form A vs. Form B) reliability coefficients (2) that were adequate on only seven or eight of the eleven factors. See Table 1. This inadequacy arises largely from some factors being measured on a bare minimum of 3 or 4 items in one form. Accordingly it is advocated that only seven or eight independent factors be routinely measured in standard clinical use and that the remaining three or four measures serve an exploratory purpose, as "located nuclei" from which further research can, by extension into new items, build up better factor scales.

Meanwhile the test has been initially standardized for every factor on a normal population of 380 student and non-student adults ranging from 18 to 68 years of age. The instructions, which are given in standard form by the voice on the record, are set out below. The I.P.A.T. Music Preference Test of Personality (3) is thus normally presented simply

as a test of musical preferences, but the implication that we were psychologically interested in the results from the standpoint of personality measurement was realized, at least by the normal group, in this particular experiment.

#### First Issues Needing Research

Now that such a measuring instrument is available, a number of researches immediately suggest themselves, especially in applied psychology. Concerning its promise as a personality test it is at once apparent from inspection of the actual musical excerpts found to be highly loaded in the various factors, that these factors are not merely culturally-determined groupings, corresponding to musical "schools" or periods (with one possible exception among the eleven factors: F 1). With this superficial interpretation rejected we may next examine the hypothesis that these factors correspond to what have been called major "hidden premises" in the logic of personal preference (1). For these hidden premises of choice decision, according to our hypothesis as stated elsewhere (1), should be temperamental and early-environment-determined dimensions of personality itself.

If this is correct, there should be some substantial correlations between these factors and the factors on the *16 Personality Factor Questionnaire* or any other measure of the primary personality factors. This at least is the hypothesis upon which the whole of the present investigation has been carried forward. If the musical choices are determined by personality factors, i.e., by emotional needs and constitutional tempers, we should expect, further, that various neurotic and psychotic syndromes, which are themselves explicable in terms of combinations of personality factors, and sometimes in terms of single personality factors, should show correlations with the musical choices. The immediately needed investigations, therefore, seem to be: (1) a study correlating the music factors with primary personality factors, in a normal group; and (2) a comparison of psychotics and normals in terms of musical preference factor profiles.

The hypotheses that the music factors cor-

respond to needs or to temperamental factors can be tested by this design, but one should also recognize that a third possibility exists—namely that the discovered music factors represent affective mood states, temporary dynamic stimulus conditions, physiological influences, etc. This alternative, however, need not be investigated unless the present search for stable personality associates proves abortive. Some “function fluctuation” associated with mood will almost certainly exist and it will attenuate our correlations. But if our hypothesis is correct that the major associations will be found in relation to relatively stable personality structures, then it could seem better to track down this residual, “fluctuation” variance later. At that point not only the associations of the music factors with mood, but also the individual tendencies to high or low fluctuation on the music factors will bring in relationships of further importance for understanding musical preference and personality.

A fourth design of research which is also immediately needed is a factorization of a population of psychotics, to see whether the *structure* of factors is the same there as in a normal group. Unless there is some fairly close resemblance of the factor structure in the two groups, it would indeed be illogical to measure psychotics on the same dimensions as those found among normals. Accordingly, we have also gathered data for factorization of the same 120 excerpts on a population of 100 psychotics, and this will be intercorrelated and factorized if statistical man-hour resources can be provided by the Music Research Foundation.

The general reaction of cultivated listeners to the above propositions has been that our hypotheses neglect the role of intellectual and cognitive functions in musical appreciation. Our argument is that these functions are not primary but are only means to ends—technical rationalizations of the aesthete, perhaps changing superficially with cultural climate—for satisfactions which are deeper and more stable. Initial experimental support for our position is given by the fact that the music factors do not apparently correspond in content to

cultural or technical dimensions. A research designed to tackle this question more positively has meanwhile been set in motion. It consists of an experiment in which fifty choices in pictorial art, thirty choices in architecture, and forty choices in sculpture are intercorrelated and also correlated with the factors in musical choice. If the same factorial dimensions appear here, aligning themselves with the music factors, and cutting across periods and cultural integrations, there will be additional evidence that we are proceeding beyond technical, cultural or historical patterns.

### Design of the Experiment

The first part of our investigation, that with normal subjects, called for the administration of the Musical Preference Test to a normal population which should be: (1) well varied in personality; and (2) simultaneously measured on a sufficiently reliable and valid measure of the primary personality factors. The main contribution to the test population consisted of 102 male and female subjects, 76 of whom were University of Illinois students, ages 18 to 29, and 26 of whom were “general adults,” ages 30 to 81. The remainder were tested in a second sub-group consisting of 55 students, both male and female, ages 17 to 28. Since we needed to apply a personality test which deals with primary and independent personality dimensions of known associations we employed the I.P.A.T. *16 Personality Factor Questionnaire*, which is also convenient for group administration with reasonably literate populations. The 16 P.F. includes intelligence as one dimension. Each of the 157 subjects, therefore, took a one-hour music preference test in which both forms A and B of the music test were administered, and a half-hour silent session in which Form A of the 16 P.F. Test was administered. The instructions in the Music Preference Test are on the beginning of the record, and are as follows:

“This is a test of your likings and dislikings in music. Your score has nothing to do with how much you agree or disagree with popular tastes, but only with how much you agree with yourself;

that is, with how consistent<sup>1</sup> you are. So try to say, as each piece is played, whether you yourself like it; whether it is pleasant, so that you would like to hear more of it, or whether you would just as soon have it switched off.

"On the score sheet before you are numbers for the fifty pieces that will be played, each for less than half a minute. As each comes to an end, underline L, I, or D, opposite that number, indicating you like it, or have an intermediate, indifferent reaction, or dislike it. Dislike does not mean that you hate it, but only that you don't particularly like that kind of music. In fact you should aim to have just as many D's as L's underlined when you get to the end. Try not to use I for intermediate more than you need. In fact, you should expect to end up with very roughly one-third L's, one-third I's and one-third D's. But don't bother about that too much. Just give your reactions as truthfully as possible. . . ."

The administration of the Music Preference Test to a group of psychotics took place at Kankakee State Hospital, Kankakee, Illinois. In this case the subjects were taken in small groups of three or four at a time, in order that it might be ascertained that they were appropriately responding on the answer sheets to every piece of music. It is well known that diagnoses in different mental hospitals do not agree very highly (as shown on the individual cases transferred from hospital to hospital), and that the very proportions of manic-depressives, schizophrenics, hysterics, and other psychotic syndromes, as diagnosed in different institutions, may vary considerably. As usual a good deal of difficulty was experienced in obtaining a sufficient sample of some psychiatric syndrome groups. In accepting the group divisions finally used the criterion for classification was naturally the hospital diagnosis as reached in case conferences. A total group of 98 psychotic patients was obtained consisting of 36 alcoholics, 22 schizophrenics of mixed types, 10 manics, 7 paranoids, and 23 of other categories each not sufficient in number for separate use in our study. The subjects were both male and female, the age range being approximately 25 to 60 years.

<sup>1</sup> This obviously asks the person to be "true to himself" and to give his considered judgment; with advanced music students on the other hand it might be interpreted as being consistent with regard to musical "schools," but our subjects were not music students.

## Results for the Normal Personalities

The findings for the normal group will first be described. Our initial interest turns on the reliabilities, a minority of which, as mentioned above, were low enough to suggest dropping certain factors. These correlations are presented first as consistency (split-half) coefficients in Table 1, Part A and secondly as coefficients of equivalence (correlation of Form A with Form B) in Part B of Table 1.

The equivalence coefficients perhaps do not do justice to the tests because the highest loaded items were in every case put in the A form, since, when psychometrists are unable to use the full length test, it is the A form that they will use. This reduces the equivalences (columns 5 and 6) below the consistency coefficients (columns 2 and 4) which more truly represent the internal consistency, and are defective—for a 10-item length of scale—only on factors 3, 9 and 10, recommended to be dropped.

The correlations between the sixteen factors of the *16 P. F. Test* and the eleven factors of the *Music Preference Test* were worked out separately for the two populations, as a mutual check. For economy of representation the values in Table 2 are blanks except where the correlations on the two samples are of the same sign and both beyond the 1% level of significance. Then a single value—the mean correlation (Fisher's *z*)—has been corrected for attenuation, by the given reliabilities of the *Music Preference* and *16 P. F. Test* measures, and recorded in Table 2.

None of the correlations is large enough to demonstrate a one-to-one relation between the music factors and the personality factors. But the set of *16 P. F. Test* factors associated with any one music factor has a psychologically consistent and compatible character among the members in every case. For example, the personality factors correlating significantly with music factor No. 1 are dominance, surgency, toughness, radicalism and self-sufficiency—all possibly related to some second-order, comprehensive factor of temperamental toughness. Furthermore (and alternatively) the relative magnitudes of the correlations are such as *could* be compatible

Table 1  
Reliability Coefficients for Factor Measurements

Factor	Part A Consistency Coefficients (Whole Group)			Part B Equivalence Coefficients (Form A with Form B)	
	Half- Length Coefficient	No. of Items in Half	Spearman-Brown Corrected to Full Length	Sample of 102 Persons	Sample of 71 Persons
1	(.71)	5	.83	.75	.64
2	(.62)	5	.77	.42	.57
3	(.06)	5	.11 (used only experimentally)	-.10	.24
4	(.41)	5	.59	.02	.19
5	(.10)	5	.18 (used only experimentally)	.11	.39
6	(.27)	5	.43	.38	.27
7	(.41)	5	.58	.15	.11
8	(.46)	5	.63	.38	.26
9	(.00)	4	.00 (used only experimentally)	.16	-.01
10	(.14)	3	.25 (used only experimentally)	.04	.11
11	(.37)	3	.55	.28	.31

with a one-to-one relationship of music and personality factors if chance experimental error and the existing spurious correlations among the factors within both the personality and the music area could be eliminated (nota-

bly by longer scales for each factor and by dropping items in one factor scale having any correlation with another factor). A test of this possible explanation must await much further work on the purification of the pres-

Table 2  
Correlations of Music Preference Factors and Personality Factors

16 P.F. Factors	Music Preference Factors										
	1	2	3	4	5	6	7	8	9	10	11
A	—	—	—	—	—	—	—	—	—	—	-.35
B	—	—	—	—	—	—	—	—	—	—	—
C	—	—	—	—	—	—	—	—	—	—	—
E	.49	—	—	—	—	-.35	-.30	—	—	—	—
F	.46	—	—	—	—	-.33	—	—	.30	—	—
G	—	—	—	—	—	—	—	-.38	—	—	—
H	—	—	.68	—	—	—	—	—	—	-.30	—
I	-.65	—	—	—	—	-.36	.34	—	—	—	—
L	.51	-.49	—	—	—	—	—	.70	-.37	—	—
M	—	—	—	.60	-.47	—	—	—	—	—	—
N	—	—	—	—	—	—	—	.41	—	—	—
O	—	—	—	—	—	—	—	-.32	—	—	—
Q <sub>1</sub>	.38	—	—	—	.60	—	—	—	—	—	—
Q <sub>2</sub>	.48	—	—	.37	-.36	.36	—	—	—	—	—
Q <sub>3</sub>	—	—	—	—	—	.38	—	—	—	—	.35
Q <sub>4</sub>	—	-.52	—	—	—	.31	—	—	—	—	—

ent factor scales. Meanwhile, however, this explanation rests on the indication that the highest correlation for a given music factor with any personality factor is also the highest correlation for the personality factor with any music factor. For example, factor 2 has its highest  $r$  with Q4, which  $r$  is also Q4's highest  $r$  with anything; factor 3's highest is with H, which is also H's highest; the factor 4 column has its highest with M, which is also the highest in the M row, and so on, with very few exceptions (notably factor 8).

As to the consistency of psychological meaning among personality factors associated with a given music factor we may mention, in addition to factor 1 above, that factor 2 correlates negatively both with paranoid tendency and nervous tension, which tendencies have been previously found associated by Darling (2); and that factor 4, which correlates essentially with M ("Unconventionality vs. Practical Concernedness"), also has some association with Q2 ("Independent Self-sufficiency"). The alternative possibility is thus indicated, as suggested above, that where a music factor does not align itself with a first-order personality factor it may prove on further research to correspond to a second-order factor uniting the personality factors in some underlying common influence. For this reason music factor 1 has been called "Tough Sociability vs. Tenderminded Indi-

viduality," which contingently restricts the meaning pretty closely to the psychological bi-polarity of personality factor I, with which it is most associated, but also suggests features of the other factors with which it has some degree of association. The over-all description of the personality dimension associated with this particular music factor thus becomes remarkably similar to the Tender-vs. Tough-minded continuum described by William James (6).

#### Results for the Abnormal Personalities

As stated above, in the account of design, the test was administered to 98 hospitalized psychotics, divided into those four major syndrome groups which had each a sufficient number of well-diagnosed cases to promise some significance of differences, if such should exist.

The means and sigmas on all 11 factors are shown for normals, for abnormals as a whole, and for the four abnormal syndrome groups, in Table 3.

The differences are examined below by the  $t$  test, first with respect to the differences between the main psychotic group and the psychotic sub-groups, on the one hand, and the normal group on the other, with results as shown in Table 4. Nothing below a 10% probability is recorded in the P column.

Table 3  
Scores of Normal and Abnormal Groups

Factor	Normals <i>n</i> = 369		Abnormals <i>n</i> = 98		Alcoholics <i>n</i> = 36		Schizophrenics (D-P) <i>n</i> = 22		Manics <i>n</i> = 10		Paranoids <i>n</i> = 7	
	Mean	Sigma	Mean	Sigma	Mean	Sigma	Mean	Sigma	Mean	Sigma	Mean	Sigma
1	13.6	5.7	14.5	4.1	15.1	3.7	13.4	3.7	15.8	4.6	12.4	6.3
2	10.7	4.1	8.7	4.7	6.4	4.4	10.0	4.7	10.2	3.6	12.1	4.6
3	9.6	2.7	8.9	2.4	9.0	2.1	8.6	2.3	9.0	1.4	9.4	1.7
4	6.8	3.3	4.8	2.4	4.1	3.0	5.5	2.5	8.2	2.6	5.4	2.1
5	12.2	2.8	11.6	2.7	12.2	2.3	11.5	3.0	12.7	2.3	10.4	1.2
6	8.4	3.1	10.8	2.8	11.4	2.7	10.7	2.8	10.2	2.2	9.4	2.6
7	8.3	3.2	7.0	2.6	5.8	2.4	7.8	2.6	7.2	1.6	7.1	2.5
8	8.0	2.6	7.9	2.2	7.3	2.1	8.4	1.9	9.1	2.6	8.4	2.7
9	7.3	3.0	9.2	2.1	9.0	2.1	9.4	2.0	9.0	2.2	8.1	1.4
10	5.6	2.1	5.9	2.0	5.5	1.9	6.0	1.7	5.9	2.1	6.6	1.5
11	6.1	2.1	5.6	2.9	6.6	2.5	5.4	3.0	3.6	2.4	3.4	2.4

texturally) music in favor of clear harmonic progressions, sweet melodies and subordinate accompaniment. The exception to this pattern is the manic group, which, on its distinguishing factor (No. 4), prefers fast, exhilarating, stimulating pieces with textural complication, rhythmic variation and less obvious melodic outlines. These associations might roughly be explained in terms of empathy, but as more evidence accumulates they should receive more direct research investigation, especially in the light of such research approaches as those of Rigg (7, 8).

### Summary

1. A previously completed factor analysis of 120 very diverse musical excerpts was used as a basis for construction of a Music Preference Test of Personality, set up to measure eleven factors by 100 items on two sides of a long-playing ( $33\frac{1}{3}$  R.P.M.) record. As the equivalence of the A and B forms is inadequate for three or four of the factors, it is recommended that these be reserved for research improvement, by item analysis, and that the remaining seven or eight factors alone be used as internally valid measures in routine applied psychology, notably in seeking external validities by predictions in clinical and guidance psychology.

2. Since the established groupings of items do not correspond to musical schools or periods (though possessed of some consistency of musical character) it is hypothesized that they represent dimensions of personality (especially of temperament) determining taste. Correlation with the *16 Personality Factor Questionnaire Test*, on normal populations of 102 and 71, confirmed this by yielding many significant correlations.

A one-to-one relation of music preference and personality factors cannot be proven by these results, since both measures of factors are imperfect. But the correlations, corrected for attenuation, are at least consistent with the hypothesis that, but for contamination, the same personality dimensions determine, in all but two cases, both the verbal and the

music preference factors. Contingent titles have been given to the music preference factors in accordance with the personality associations. These titles proceed on the probability that most music factors are primary personality factors though some may be second-order personality factors.

3. Application of the Music Preference Test to 98 patients in mental hospitals revealed several factor measure differences, significant at the 1% level, between psychotics and normals and between various psychotic syndrome groups. If confirmed on further samples, these pattern differences are so marked as to make the test a valuable adjunct to psychiatric diagnosis. The meaning of the music factors as indicated by the personality factor correlations agrees well with the meaning as found independently in terms of the associations with psychotic syndrome groups. These scales might therefore have value in throwing further light on individual psychotic syndromes.

Received February 24, 1953.

### References

1. Anderson, H. H. and Anderson, G. H. *Projective techniques*. Chap. 2. New York: Prentice Hall, 1951.
2. Cattell, R. B. *A guide to mental testing*. London: University of London Press, Third Edition, 1953.
3. Cattell, R. B. and Anderson, J. C. *The I.P.A.T. Music Preference Test of Personality*. The Institute for Personality and Ability Testing, 1608 Coronado Drive, Champaign, Illinois, 1953.
4. Cattell, R. B. and Saunders, D. R. Musical preferences and personality diagnosis. I. A factor analysis of 120 themes. *J. gen. Psychol.*, 1953, in press.
5. Cattell, R. B. and Wenig, P. W. Dynamic and cognitive factors controlling misperception. *J. abnorm. soc. Psychol.*, 1952, 47, 797-809.
6. James, W. *Pragmatism: a new name for some old ways of thinking*. London: Longmans, 1911.
7. Rigg, M. G. Musical expression: an investigation of the theories of Erich Sorantin. *J. exp. Psychol.*, 1937, 4, 442-455.
8. Rigg, M. G. Speed as a determiner of musical mood. *J. exp. Psychol.*, 1940, 5, 566-571.

## A Rating-Scoring Method for Free-Response Data

Ralph R. Canter, Jr.

*University of California, Berkeley*

In a study designed to evaluate a human relations training program for executives and supervisors (reported in detail elsewhere, 1) a forced-normalizing method was employed to score written answers to open-ended questions. Answers to four questions contained in a specially prepared *Supervisory Questionnaire* were evaluated by four raters in accordance with the procedures outlined in this paper. This questionnaire was one of a number of tests and other questionnaires administered to an experimental (trainee) group and a matched control group. The N was 18 in each group.

The questions used in the *Supervisory Questionnaire* were developed in light of three considerations: (a) the trainees should be given an opportunity to express themselves in their own words concerning the kinds of problems they had in their jobs; (b) the questions should not be drawn from the course content, but should be directed toward the individual supervisor in his job; and (c) the questions should not be structured in such a fashion that certain kinds of answers would seem important. As an example, one question used was: "Do you feel you have the kind of cooperation from your employees that you want? What do you think accounts for this?"

### Rationale and Procedure<sup>1</sup>

It was hypothesized that if we were to take all the written answers to a single question made by the persons in both the experimental (E) and control (C) groups and have raters sort them into *n* categories, the E and C pretest distributions should be almost identical and have the same mean. However, it was thought the E responses following training

would be distributed by the raters in such a manner that the E mean would be reliably higher than the C mean, thus enabling us to conclude that the training was effective in producing change along the dimensions of the questions used.<sup>2</sup> The forced normal distribution of judgments was considered as an effective method to use in this situation. The derived procedures will be described in the context of the investigation.

Since there was a total of 36 E and C responses to each question, a normal area distribution with an N of 36 was determined for seven categories, this number being arbitrarily used because of the small N and the relative ease for the raters. The numbers of cases falling in the respective categories were: 1, 3, 8, 12, 8, 3, and 1.

Four raters were used, all being social scientists experienced in dealing with written questionnaire item responses and having specific knowledge of desirable supervisory practices and qualities. Each was given written instructions, a summary of which follows. The general nature of the task was described and the specific questions were listed. The rater was asked to judge the responses to these questions in terms of the degree to which they reflected over-all supervisory quality. The rater was told that the responses came from Ss in E and C groups, but that the procedure required him to be in ignorance of whether a respondent was in the E or C group. He was then instructed to sort the responses to the first question into the seven categories in accordance with the assigned numbers (i.e., 12 responses in Category No. 4, 8 each in Category No. 5 and No. 3, and so on). The exact procedure was described, essentially involving

<sup>1</sup> The writer is indebted to Dr. F. M. Fletcher of Ohio State University for assistance in developing this method. Thanks for services as raters are accorded to Dr. J. H. Hemphill and Dr. M. S. Seeman of the Personnel Research Board, Ohio State University, and Dr. E. E. Ghiselli and Mr. Richard Barthol, University of California.

<sup>2</sup> In the major study (1) the trained supervisors were found to have gained in mean score at a statistically significant level of confidence over the untrained supervisors on the *Supervisory Questionnaire*. This measure also intercorrelated highly with other tests and measures on which statistically significant gains were found.

separation of the best and poorest responses at each step. He next proceeded to the second question and so on. The rater was never informed as to whether he was dealing with a set of pretest or posttest responses.

Each response was scored by summing the category values assigned by the four raters. The number of a category was used as a score. For example, the four judges may have respectively placed a response in the following numbered categories: 2, 3, 3, and 4. The response would receive a score of 12. Each individual's four scores were then summed, this being his over-all questionnaire score (which was treated statistically within the framework of the larger investigation).

Records by separate item scores and by total scores were kept for each rater so that inter-rater reliability could be estimated.

Inter-Rater Reliability

Table 1 contains the Pearsonian correlation coefficient between raters on the summated questionnaire score (i.e., total of scores assigned by each rater to each respondent on

each of the four items) for both the pretest and posttest. The inter-rater correlations are not reported for the four separate items; we wish only to note that the range of these correlations on the pretest was from 0.31 to 0.71 with a mean of 0.47, and 0.26 to 0.79 on the posttest with a mean of 0.49.

Discussion

The inter-rater reliabilities appear to be quite adequate and within the usual range of reported reliabilities. Rating each question separately has the effect equivalent to adding more raters (2). Also, it is possible to assume that a fairly high degree of unidimensionality is accorded to each item (the rater has only a single question to keep before him). The responses can be viewed as homogeneous since the rater has only a single criterion to keep in mind—in this study goodness of response as related to supervisory quality. These conditions act to increase reliability.

In using a technique such as this it appears that some of the hazards involved in trying to get scales can be avoided. However, Suchman (4) has pointed out the difficulties with "non-itemized" judgments or ratings, noting especially that such procedures produce no definition of the variable under consideration. With this we must concur. But much depends upon the uses to be made of such ratings. In the example used the major intent was to determine whether the training appeared to have any effect at all on the trainees' free responses about how they performed in their jobs.

Subsequent studies would be required to specify the correlations between the particular training content and the observed effects, as usually is the case. From this standpoint the method proposed here is best viewed as one which determines whether further studies are warranted.

Summary

A rating-scoring technique for evaluating free response answers was developed through the use of a forced normal distribution of judgments. An example of its use in a study

Table 1

Inter-Rater Reliability Coefficients for Supervisory Questionnaire Summated Scores

Pretest			
Rater	A	B	C
B	.52		
C	.52	.63	
D	.54	.63	.66
Average Intercorrelation Coefficient*			.58
Total Summated Questionnaire Score Reliability (Corrected by Spearman-Brown formula with four raters)			.85
Posttest			
Rater	A	B	C
B	.71		
C	.69	.68	
D	.68	.70	.51
Average Intercorrelation Coefficient*			.65
Total Summated Questionnaire Score Reliability (Corrected by Spearman-Brown formula with four raters)			.88

\* Obtained by formula 118, p. 197, Peters and Van Voorhis (3).

evaluating a human relations training course was described wherein the criterion used by four judges was over-all supervisory quality as revealed in pretest and posttest written responses made by experimental and control subjects in regard to their job performance. Satisfactory inter-rater reliabilities were found.

*Received February 2, 1953.*

### References

1. Canter, Ralph R. An experimental study of a human relations training program. *J. appl. Psychol.*, 1951, 35, 38-45.
2. Furfey, P. H. An improved rating scale technique. *J. educ. Psychol.*, 1926, 17, 45-48.
3. Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
4. Suchman, E. A. The logic of scale construction. *Educ. psychol. Measmt.*, 1950, 10, 79-93.

## Factors Affecting Student Evaluation of College Faculty Members

Alexis M. Anikeeff

*Department of Institutional and Industrial Management, Mississippi State College<sup>1</sup>*

For the purpose of evaluating the potential usefulness of a student evaluation program of college faculty members, an investigation was initiated to determine the effect of grading leniency upon merit rating scores of faculty members in the School of Business and Industry. To the extent that grading leniency was highly correlated with merit rating scores, the usefulness of student evaluation could be seriously questioned.

As a corollary of the basic study, the relationship between absence extensiveness and faculty ratings was investigated to supplement data uncovered in a previous study about the effectiveness of an unlimited absence regulation. To the extent that highly rated instructors attract students to their classrooms despite an unlimited absence regulation, restricting the number of absences which students are allowed may merely serve to bolster the security feelings and egos of lowly rated instructors, and further frustrate the students who are engaged in the program of evaluating faculty members.

### Procedure

Grading leniency was determined by deriving the arithmetic mean of quality points issued by each faculty member to his students, and then, by ranking each faculty member according to obtained means. Owing to the selection process operating during a four year college curriculum, the average class grade increases progressively from the freshman to the senior level. Consequently, in order to control variation based upon academic level, rather than upon faculty leniency, separate ranking distributions were made for freshman-sophomore and junior-senior levels.

The determination of absence extensiveness was accomplished by ranking nineteen faculty members according to the median number of

class absences accumulated in their classes. Ranking distributions were made for freshman-sophomore and junior-senior levels, as well as the freshman through senior levels combined.

Since the evaluation data of faculty members were available in ranked form, Spearman  $\rho$  was used to determine the relationship between grading leniency and merit rating scores. The same technique was used to uncover the relationship between absence extensiveness and merit rating scores. In addition, multiple correlation analysis was used to determine the combined effect of grading leniency and absence extensiveness upon the merit rating scores of faculty members.

Absence and grading data are based upon reports which 19 faculty members submitted to the Dean of the School of Business and Industry. Faculty evaluation data are based upon information secured during regular class periods by members of an honorary scholastic fraternity after mid-semester grades were posted, but before final grades were assigned. Instructors were rated on an eight-point graphic rating scale which permitted distribution of judgments along a continuum of five verbally described points. The following factors were used: 1. knowledge of the subject; 2. class preparation; 3. clarity of speech; 4. avoidance of sarcasm; 5. fairness in grading; 6. absence of mannerisms; 7. creation of interest in subject matter; and 8. ability to control temper.

Under provisions of the evaluation program, an instructor left his classroom earlier than usual and permitted students to rate him during his absence. Whenever an instructor was rated by less than 50 students, or by less than three different classes, he was excluded from the standardizing population, and is also excluded from the present study. The obtained ranking distributions are based upon approximately 1,500 cases.

<sup>1</sup> Now at Oklahoma A & M College.

## Results

On the freshman-sophomore level, a very significant and moderately high positive relationship is found between grading leniency and faculty merit rating scores. Consult Table 1 for more specific information. Through the use of a coefficient of determination described by Guilford,<sup>2</sup> it is evident that approximately 53% of the variance in the rating received by an instructor who teaches freshman-sophomore level courses can be attributed to grading leniency. No significant relationship is found between obtained rating and grading leniency on the junior-senior level. On a combined basis, freshman through senior, a significant but moderate relationship exists between obtained rating and grading leniency. Approximately 25% of the variance in an instructor's rating can be attributed to grading leniency on an over-all four-year basis.

Table 1

Spearman Rank-Difference Correlations for Faculty Members Ranked on Three Factors

Class Level	N	Compared Rankings		
		Grades vs. Absences	Rating vs. Absences	Grades vs. Rating
Freshman-Sophomore	13	-.21	-.17	.73**
Junior-Senior	17	-.19	-.26	.43
Freshman-Senior	19	-.08	-.53*	.51*

\* Significant at the five per cent level.

\*\* Significant at the one per cent level.

No significant relationships are found between obtained rating and absence extensiveness, on either the freshman-sophomore or the junior-senior levels. However, a significant, negative, and moderate relationship is found between instructors ranked according to merit rating scores obtained by student evaluation and the same instructors ranked according to the median number of absences accumulated in their classes, on a four-year breakdown. Thus, an instructor with a lower number of absences receives the higher rating. Approximately 28% of the variance in absence ex-

<sup>2</sup> Guilford, J. P. *Fundamental statistics in psychology and education*. (2nd Ed.) New York: McGraw-Hill, 1950.

Table 2

Multiple Correlations Between Faculty Ratings and Grading Leniency Combined with Absence Extensiveness

Class Level	Correlation Data		
	N	R	SE <sub>R</sub>
Freshman-Sophomore	13	.73**	.15
Junior-Senior	17	.47	.21
Freshman-Senior	19	.70**	.13

\*\* Significant at the one per cent level.

tensiveness can be accounted for by the merit rating scores of instructors. By the same token, 28% of the variance in an instructor's rating can be attributed to absence extensiveness.

No significant relationship is found between instructors ranked according to grading leniency and instructors ranked according to the median number of class absences accumulated in their classes. The lack of such relationship is evident on the freshman-sophomore, junior-senior, and the combined freshman through senior levels.

When absence extensiveness and grading leniency are combined, and the overlap between these factors is held constant, the combination of these factors accounts for approximately 53% of the variance in an instructor's rating on the freshman-sophomore level. Consult Table 2 for detailed data. The same combination of factors accounts for about 22% of the variance in an instructor's rating on the junior-senior level, and for approximately 50% of an instructor's rating variance on the freshman through senior levels.

## Discussion

For the sample used, the grades which faculty members assign students are reflected in the quality of rating which students assign faculty members. The extent to which students evaluate faculty members according to class grades received varies with academic levels of the students. Grading leniency accounts for almost three times as much variance in faculty ratings on the freshman-sophomore level, as it does on the junior-senior

level where the relationship is not statistically significant. Conceivably, students who survived the selection process operating during the first two years consider faculty grading leniency as a relatively unimportant criterion on which to base evaluation of faculty members.

Class absences are negatively correlated with faculty ratings. These results may indicate varying degrees of student interest in the classroom behavior of faculty members. If such is the case, lowly rated faculty members apparently repel students from their classrooms, and accordingly, accumulate disproportionate numbers of class absences. However, in the interpretation of the relationship between absence extensiveness and faculty rating scores, it should be noted that the factor of absence permissiveness remains uncontrolled.

Absence permissiveness could operate directly or indirectly. Direct operation could involve open avowal or subtle implication, on the part of highly rated instructors, that class attendance is not required for satisfactory course performance. Conversely, lowly rated instructors may insist upon daily attendance to the point where daily recitation grades carry an unduly preponderant weight in the determination of final class grades. Loading course examinations with textbook questions, while minimizing the inclusion of lecture ques-

tions, could illustrate the manner in which indirect absence permissiveness would operate.

### Summary

Nineteen faculty members were ranked in accordance with the merit rating scores assigned to them by their students. Using Spearman *rho*, merit rating ranks were correlated with grading leniency and absence extensiveness rankings of the same instructors.

1. Grading leniency correlated highest with merit rating scores on the freshman-sophomore level, and lowest on the junior-senior level.

2. Absence extensiveness correlated negatively on all academic levels, but the correlation was significant only on the combined four-year breakdown.

3. The selection process operating during the freshman-sophomore years could reasonably account for a low and statistically non-significant relationship between grading leniency and student evaluated faculty merit ranking on the junior-senior level.

4. Class interest of students could account for the negative relationships between faculty members ranked according to the number of class absences found in their classes and the same instructors ranked according to teaching competence as evaluated by students.

*Received March 6, 1953.*

## Estimating Grade Reliability<sup>1</sup>

Scarvia B. Anderson

*Naval Research Laboratory, Washington 25, D. C.*

When grade point ratio is used as the criterion of school "success" or "failure," the need for an adequate estimate of the reliability of the ratios presents a recurring problem. The author encountered the problem most recently in a study at George Peabody College for Teachers of the value of certain entrance tests in predicting freshman grade point ratio (1). If the results were to be used for the selection and counseling of students and for the determination of an adequate testing program, some estimate of the reliability of the criterion seemed essential.

The tests used were the American Council on Education Psychological Examination; the Cooperative Reading Comprehension Test; the Cooperative Mechanics of Expression Test; and Otis Quick-Scoring Mental Ability Tests for grades four through nine, which were used as practice tests. Multiple correlations of test scores with weighted grade point ratios were .62 and .59 for one quarter and three quarters, respectively. It was reasonable to assume that grade point ratios for three quarters were more reliable than those for one quarter, but a statistical estimate of such reliability seemed desirable in the final interpretation of the results.

Ebel (4) has presented a method, based upon analysis of variance, for estimating the reliability of sets of ratings, and in addition has considered in some detail the relationship between this procedure and those of Horst (5), Snedecor (7), Clark (2), Peters (6), and Cureton (3). He concludes that the intra-class formula, such as his, Cureton's, and Snedecor's, is generally preferable to the average intercorrelation or generalized re-

liability formula, such as Horst's, Clark's, and Peters'.

In the present paper, we shall discuss: (a) the differences obtained when the Horst and Cureton formulas were used to estimate the reliability of the *same* sets of freshman grades; and (b) a second problem, which arose in the application of the formulas, the use of unweighted versus weighted grade point ratios.

Both Cureton's formula (which, incidentally, is the result of a derivation parallel to Ebel's) and Horst's are based on the well-known generalized formula for the reliability coefficient:

$$r = 1 - \frac{\sigma_e^2}{\sigma_0^2}$$

In application here, the error variance is an estimate of the error variance of the individual means, and the observed variance is the variance of the means for all of the individuals.

Cureton's and Horst's formulas are shown in Table 1. The chief statistical differences between them may be summarized as follows:

1. Cureton uses a weighted variance for the estimate of the error variance, and Horst does not.

2. Cureton uses a weighted variance of the person means, and Horst does not.

3. Cureton divides by  $N - 1$  in the variance of the means, and Horst divides by  $N$ .

A careful study of the two methods and their respective relevance to freshman grade point ratio suggested that Cureton's technique was more appropriate for our use. If our freshmen were considered a sample of a universe of Peabody freshmen, the relevance of dividing by  $N - 1$  was apparent. In addition, we agreed with Cureton that his formula would give a somewhat better reliability estimate, since the values he uses for error variance and total variance of the person means are "statistically independent in the sense of

<sup>1</sup> The author is deeply indebted to Dr. Julian C. Stanley, University of Wisconsin, for material help in the preparation of this article and to Dr. E. E. Cureton, University of Tennessee, and Clarence W. Spence, George Peabody College for Teachers.

The opinions and assertions contained herein are the private ones of the writer and are not to be construed as official or reflecting the views of the Navy Department or the naval service as a whole.

Table 1  
Formulas for Estimating Reliability of Means of Unequal Numbers of Scores\*

Horst (5)	Cureton (3)
$\sigma_i^2 = \frac{\sum \frac{Sx_i^2}{n_i(n_i - 1)}}{N} \quad (1H)$	$\sigma_i^2 = \frac{\sum Sx_i^2}{n(\sum n_i - N)} \quad (1C)$
$\sigma_{M_i}^2 = \frac{\sum (M_i - M)^2}{N} \quad (2H)$	$\sigma_{M_i}^2 = \frac{\sum n_i (M_i - M)^2}{n(N - 1)} \quad (2C)$
$r = 1 - \frac{\sum \frac{Sx_i^2}{n_i(n_i - 1)}}{(M_i - M)^2} \quad (3H)$	$r = 1 - \frac{\frac{\sum Sx_i^2}{\sum n_i - N}}{\frac{\sum n_i (M_i - M)^2}{N - 1}} \quad (3C)$
$r = 1 - \frac{\sum \frac{SX_i^2}{n_i(n_i - 1)} - \sum \frac{M_i^2}{(n_i - 1)}}{\sum (M_i - M)^2} \quad (4H)$	$r = 1 - \left( \frac{N - 1}{\sum n_i - N} \right) \left( \frac{\sum SX_i^2 - M \sum SX_i}{\sum (M_i SX_i) - M \sum SX_i} - 1 \right) \quad (4C)$

\*  $N$  = number of individuals,  $Sx_i$  = sum of the deviation scores for individual  $i$ ,  $SX_i$  = sum of the raw scores for individual  $i$ ,  $n_i$  = number of scores for individual  $i$ ,  $n$  = mean number of scores for  $N$  individuals,  $M_i$  = mean score for individual  $i$ ,  $M$  = mean score for  $N$  individuals.

analysis of variance, while those given [by Horst] . . . are not quite independent" (3, p. 2). Still more important, however, since the results of the freshman test study were to be used for *predictive* purposes, the weighting of the variances, so that a mean based on a smaller number of measures would receive a smaller weight than a mean based on a larger number of measures, should furnish a better population estimate of reliability.

However, it was decided to use both methods in estimating the reliability of the freshman grades for one quarter and for three quarters, in order that the results might be compared for discrepancies.<sup>2</sup>

At this point, it seems well to consider briefly the first problem encountered in application of the two formulas. In the original analysis of the relationship between grades and test scores, grades were weighted according to the number of quarter hours that a

course carried:

$$M_{iwt} = \frac{\sum_{i=1}^{n_i} (h_{ci} p_{ci})}{\sum_1 h_{ci}}$$

where

- $M_{iwt}$  = weighted grade point ratio for individual  $i$ .
- $h_{ci}$  = number of quarter hours in any one course  $c$  that individual  $i$  takes.
- $p_{ci}$  = points assigned to a grade in any one course  $c$  that individual  $i$  takes.<sup>3</sup>
- $n_i$  = number of courses that individual  $i$  takes.

It was reasoned that a grade in a five hour course would be considerably more reliable than a grade in, say, a two hour course, for in a five hour course the instructor would meet with a student more frequently, would probably give more quizzes, and would generally be in a better position to give a more (subjectively) reliable grade. However, if this

<sup>3</sup> Points were assigned on the basis of the following scale: A + = 12 points, A = 11 points, A - = 10 points, B + = 9 points, B = 8 points, B - = 7 points, C + = 6 points, C = 5 points, C - = 4 points, D + = 3 points, D = 2 points, D - = 1 point, F = 0 points.

<sup>2</sup> It is realized that any technique for estimating reliability assumes independence of measures. In the case of grade point ratio, it is difficult, if not impossible, to meet this assumption. In addition to the possibility of teachers discussing among themselves the ratings they give students, a single grade point ratio may contain grades from two or more courses under one professor. In order to obtain a statistical estimate of reliability, however, one seems to have no choice but to use one of the standard reliability formulas, recognizing the limitations imposed by the conditions usually surrounding college grading.

viewpoint was maintained, several adjustments would have to be made in order to use Cureton's or Horst's formula. We could not let  $n_i$  equal the number of quarter hours taken by an individual, since that interpretation would not be compatible with the original meaning of  $n_i$  in the formulas. A letter from Dr. Cureton granted that a grade in a five hour course would probably be somewhat more reliable than a grade in a two hour course; however, he indicated that the difference would be less than might be suggested by the weights of 5 and 2, for in most courses, regardless of the number of hours, a final examination is given and instructors generally try to administer enough tests to give (subjectively) reliable grades.

Unweighted grade point ratios were computed as follows:

$$M_i = \frac{\sum_1^{n_i} P_{ci}}{n_i},$$

where  $M_i$  = unweighted grade point ratio for individual  $i$ .

These correlations between weighted and unweighted grade point ratios were obtained:

- .98 between weighted and unweighted grade point ratios for three quarters.
- .96 between weighted and unweighted grade point ratios for one quarter.

As a result, any advantage of weighting seemed so small that we were satisfied to go ahead with the calculation of the reliability coefficients, using unweighted grade point ratios and letting  $n_i$  equal the number of courses that individual  $i$  took.

The Cureton raw-score formula (4C) was used first, and then since  $M_i$ ,  $M$ , and  $SX_i$  were already known, the raw score derivation (4H) of the Horst formula was used. The resulting reliability coefficients were as follows:

	Horst	Cureton
Grades for 3 quarters	.90	.90
1 quarter	.48	.63

Although the first two  $r$ 's are identical, there is a rather wide discrepancy between the last two  $r$ 's. It seems that there is logically

no statistical method for testing the significance of the difference between these two estimates of the reliability of the *same* measures, and one must return to the original formulas for an explanation of the numerical differences. The fact that the reliability coefficients for three quarter grade point ratios are the same and for one quarter grade point ratios are different is directly attributable to the weighting process used in Cureton's formula. With the large values of  $\sum n_i (M_{n_i})$  equals 13.48) for three quarters, the weighting seems to have a negligible effect on the value of  $r$ ; while with the smaller  $\sum n_i (M_{n_i})$  equals 4.62) for one quarter, the weighting process results in a considerable difference between the variance estimates substituted in Cureton's and those substituted in Horst's formula. Cureton's formula gave for one quarter a considerably smaller estimate of error variance than did Horst's.

The immediately obvious conclusions that were drawn from the computed reliability coefficients were that: (1) the use of one quarter's grades alone would not be adequate for our purposes; and (2) three-quarter grade point ratios represented a fairly reliable criterion.<sup>4</sup> The difference obtained between the Horst and Cureton coefficients for one quarter did not affect these conclusions. If there had been a question of interpretation, we should have used the reliability estimate given by Cureton's formula for the reasons already given.

In other cases where reliability estimates of unequal numbers of ratings were to be made, we would generally tend to use Cureton's formula when we were interested in reliability for the *prediction* of population behavior from a sample of that population or

<sup>4</sup> Interpretation is much more difficult than this statement indicates, however. Since many of the students had fewer different teachers than quarter-courses during the three quarters (especially in the required English sequence), grades received by an individual from quarter to quarter were by no means independent of each other. How much less the estimated reliability coefficient of .90 would be if true independence existed cannot be judged from these data. It is interesting to note that despite markedly poorer reliability of first-quarter grade point ratios, the multiple  $R$  between test scores and first-quarter GPR's was slightly higher than for the entire year (.62 vs. .59).

Table 2

Means and Standard Deviations for Test Scores  
and Course Grades

Variable	Mean	Standard Deviation
ACE Q	41.7	7.9
ACE L	61.2	16.3
Cooperative English	141.2	29.5
Cooperative Algebra	34.0	12.2
Minnesota Paper Form Board	45.3	7.9
Bennett Mechanical	32.7	12.6
Mathematics	6.3 <sup>a</sup>	2.8
English	6.1 <sup>a</sup>	2.5
Engineering Drawing	7.4 <sup>a</sup>	2.0
Civil Engineering	2.9	1.2
Mechanical Engineering	2.4	1.2
Engineering Problems	2.7	1.3

<sup>a</sup> Summation of three grades.

fall quarter, 1950, through the fall quarter, 1951, constituted the criteria for the study. Freshman year grades generally take care of most of the screening of engineering candidates at the University of Tennessee, as failures are much more unlikely after the first few quarters in the engineering curriculum. The entering class is a relatively heterogeneous group, as no selection procedures are used other than a minimum mathematics requirement of four high school units.

Instead of using the mean point hour ratio for all courses combined, correlation coefficients were computed for the various tests with grades in the courses in the freshman engineering curriculum. Table 1 shows the various courses and tests for which correlations were computed.

## Discussion of Results

Though the coefficients found in Table 1 are not especially high, several of them are sufficiently so to be regarded as useful for selection or guidance situations. With a population consisting of high school students, a higher correlation would be hypothesized for this more heterogeneous group.

The best predictive instrument in the battery used seems to be the Cooperative Algebra Test; the Cooperative English Test ranks second. The Bennett Mechanical Comprehension Test tends to get better correlations with grades than either of the A.C.E. scores. In an unpublished master's thesis at the University of Tennessee, Tarvin (6) found that the algebra and English tests yielded higher correlations than either A.C.E. score among freshman students. From these data and other studies (4, 11), the predictive value of so-called scholastic aptitude tests such as the A.C.E. must be questioned in comparison to outright achievement tests.

Further examination of Table 1 will reveal that in different courses different instruments may be the most effective predictors. It is no surprise to find the algebra test predicting mathematics grades best, and the English test performing in a similar fashion for English grades. The Bennett is clearly the best predictor in engineering drawing instead of the Minnesota Paper Form Board as might have been expected. No test emerges as a good predictor for civil engineering. This may reflect to some extent the unreliability of grades in this course though further evidence is needed. In mechanical engineering the Eng-

Table 3  
Multiple Correlation Work Sheet, Test Scores and Course Grades,  
University of Tennessee Engineering Freshmen

Criterion	Test Predictors	R	N
Mathematics	Alg. (.558) + Bennett (.612)** + Eng. (.622)	.622	82
English	Eng. (.608) + Alg. (.614) + Bennett (.622)	.622	86
Engr. Drawing	Bennett (.453) + Alg. (.505)* + ACE L (.541)*	.541	80
Civil Engr.	Alg. (.290) + Eng. (.310) + Minn. (.314)	.314	62
Mech. Engr.	Bennett (.556) + Eng. (.664)** + Alg. (.686) + ACE Q	.708	58
Engr. Problems	Alg. (.496) + Bennett (.595)** + ACE Q (.612)	.612	59

\*\* Increment in R significant at 1% level.

\* Increment in R significant at 5% level.

lish test stands out as the best predictor. Does this reflect an emphasis in grading in this course on competence in English usage? The algebra test and the Q score seem to be the best predictors in engineering problems, though the Bennett provides a moderate correlation coefficient.

It is interesting to note that the Q score is more valuable than the L for this engineering group in the courses considered. This, of course, is contrary to the usual findings with the A.C.E. in other curricula (3, 11). The Minnesota Paper Form Board yielded generally the lowest correlation coefficients of any of the tests.

Multiple correlations were then computed for four of the criterion variables, grades in English, engineering drawing, engineering problems, and mathematics. Table 3 presents these data showing the best multiple correlations that can be obtained with the tests used.

The addition of further tests does not add much in the case of English and mathematics where the zero order correlations were moderately high in the first place. In engineering drawing and engineering problems the extra tests appreciably contribute in improving the correlation coefficients, from .496 to .612 for the problems course, and from .453 to .541 for the drawing course. Additional tests seem warranted for more reliable prediction in the case of these two courses.

#### Summary

In conclusion, it can be stated that this study has demonstrated the satisfactory applicability of several economical (in terms of administration and cost) tests for the problem of selecting or guiding prospective engineering students. The tests which produced

the best correlation coefficients were the Cooperative Algebra, Cooperative English, and the Bennett Mechanical Comprehension Tests. The A.C.E. Psychological Examination and the Minnesota Paper Form Board were not as adequate. The findings in this study seem to generally confirm those of previous investigators.

Received February 11, 1953.

#### References

1. Berdie, R. F. Differential aptitude tests as predictors in engineering training. *J. educ. Psychol.*, 1951, 42, 114-123.
2. Berdie, R., Dressel, P. and Kelso, P. Relative validity of the Q and L scores of the ACE Psychological Examination. *Educ. psychol. Measmt.*, 1951, 11, 803-812.
3. Cole, A. W. *Predicting success in engineering*. Department of Vocational Education, University of Arkansas: Fayetteville, Arkansas, 1951.
4. Hellmer, L. A. Unpublished data from the University of Illinois, 1953.
5. Johnson, A. P. College Entrance Examination Board Mathematical Tests: (a) and the Pre-Engineering Inventory; and (b) as predictors of scholastic success in colleges of engineering. *Amer. Psychologist*, 1950, 5, 353. (Abstract)
6. McNemar, Q. *Psychological statistics*. New York: Wiley, 1949.
7. Moore, J. E. A decade of attempts to predict scholastic success in engineering schools. *Occupations*, 1949, 27, 92-96.
8. Pierson, G. A. and Jex, Frank B. Using the Cooperative General Achievement Tests to predict success in engineering. *Educ. psychol. Measmt.*, 1951, 11, 397-402.
9. Stuit, D. B. *Predicting success in professional schools*. Washington, D. C.: American Council on Education, 1949.
10. Tarvin, J. C. *Prediction of freshman course grades at The University of Tennessee*. Unpublished M.A. Thesis, University of Tennessee, 1951.
11. Wallace, W. L. The prediction of grades in specific college courses. *J. educ. Res.*, 1951, 44, 587-597.

## Academic Achievement in Engineering Related to Selection Procedures and Interests

Louis Long and James D. Perry

*Division of Testing and Guidance, The City College of New York*

Over the past ten years the matter of improving the selection of engineering students has been under investigation at the City College. A variety of tests has been used in conjunction with the high school average in determining which students should be admitted. The ACE Psychological Examination, the Cooperative General Achievement tests, and the Pre-Engineering Inventory have been used as well as some special tests constructed for the college by Kenneth W. Vaughn. Tests have been added and eliminated on the basis of studies relating test scores to academic grades. At this point the program has become fairly stable and consequently it was felt that a report based on the present battery might be of interest to other colleges.

Specifically this study was designed with two purposes in mind: to evaluate the effectiveness of high school averages and scores on entrance examinations as a basis for predicting four-year grade-point average in an engineering college, and to study the relationship between the four-year college grade-point average and ratings on two standard interest questionnaires (the Strong Vocational Interest Blank and the Kuder Preference Record). It was also hoped that ratings on the Kuder obtained from students during their freshman year might be compared with those obtained during their senior year. The number of students taking the questionnaire on both occasions was small, but since there are not much data of this type available we shall, nevertheless, present them.

The results reported are based on students graduating from the School of Technology of the City College during the calendar year of 1951. Of the 521 graduates, 433 were included in one or more phases of this study. Since all available data were used in each part of the study the number of cases varies from one part of the study to another.

A four-year college average (weighted according to credits and grades), calculated by the School of Technology,<sup>1</sup> was available for each graduate.

### Effectiveness of Selection Techniques

Studies based on data from previous entering classes have indicated that first term grades at City College can be most effectively predicted by using a Composite Score based upon high school average (weight of five), and scores on the following tests: Scientific Verbal Ability (weight of one), Comprehension of Scientific Materials (weight of two), and General Mathematical Ability (weight of two).<sup>2</sup>

The intercorrelations between the four-year college average, high school average, and the scores on the three tests entering into the Composite Score are presented in Table 1 along with the means and standard deviations for these variables. The correlations between the four-year college average and the other variables range from 0.30 to 0.50.<sup>3</sup> The correlation between the Composite Score and the four-year college average is 0.53,<sup>4</sup> which is

<sup>1</sup> The authors would like to take this opportunity to thank Professor John R. White for making this material available to us.

<sup>2</sup> These weights were determined by means of regression equations in which the effectiveness of these three tests as well as the following were determined: ACE Psychological Examination, General Verbal Ability, Social Science Verbal Ability, and Spatial Visualizing Ability. All of the tests, except the ACE, are part of the Inventory of Scholastic Ability and were developed by Kenneth W. Vaughn. Several of the tests are similar to those included by Vaughn in the original Pre-Engineering Inventory (13). For a detailed description of the entrance examination program at the City College the reader is referred to an article by Long and Perry (5).

<sup>3</sup> It may be of interest to the reader to compare these correlations with those reported in the literature for other colleges [2, 6; see summaries by Kandel (4), Moore (8), Stuit (11)].

<sup>4</sup> The range of the test scores has been reduced by about 10 per cent due to the elimination of students over a four-year period. The effectiveness of the tests is thereby reduced. In previous studies cor-

Table 1

Intercorrelations, Means, S.D.s: Four-Year College Average, High School Average, and Three of the Entrance Examinations (N = 182)

	Variables				
	1	2	3	4	5
1. Four-year college average	—				
2. High school average	.40	—			
3. General Math. Ability	.50	.34	—		
4. Comp. Sci. Materials	.41	.28	.61	—	
5. Sci. Verbal Ability	.30	.29	.41	.58	—
Mean	80.6	83.5	51.2	56.4	57.5
S.D.	4.8	3.9	14.1	11.1	12.6

only a slight increase over the correlation of 0.50 between the score on the test measuring General Mathematical Ability and the four-year college average, but a sizable increase over the correlation of 0.40 between high school average and the four-year college average.<sup>5</sup> Adding the other tests given as part of the Entrance Examination (see footnote 2) to the Composite Score does not bring about a significant increase in this correlation of 0.53 between the Composite Score and the four-year college average.

#### Interests and Grade-Point Average

*Strong Vocational Interest Blank.* The Strong was administered to 158 of the students as seniors (12 in Chemical, 38 in Civil, 67 in Electrical, and 41 in Mechanical Engineering). The mean standard score and the letter grade equivalents on the scale for the engineers and on the group scales are presented in Table 2. The composite profile for these students shows a B+ on the scale for the engineers and an A on the Group II scale

relations ranging from 0.55 to 0.70 have been found between the Composite Score and the first term average (5).

<sup>5</sup> It should be mentioned that about three quarters of these students were admitted on the basis of high school average alone, while a quarter were admitted on the basis of the Composite Score. This means that the test in mathematics was part of the selective procedure in only a small part of the sample, whereas the high school average was used in all instances (either alone or in combination with test scores). This situation explains to some extent why the correlation between the mathematics test and college grades is higher than that between high school averages and college grades.

(chemist, engineer, mathematician, and physicist). The correlations between the various scales of the Strong and the four-year college averages are low and not significant (see Table 2). These correlations are, of course, lowered to some extent by the fact that the academically weaker students have dropped out of engineering, as have many of those students with little or no interest in engineering.

It was interesting to note in analyzing the interest pattern of the engineering students that 82 per cent of them obtained A or B+ ratings on the Group II scale, whereas Strong reports 77.5 per cent of his criterion group for Group II obtained A or B+ ratings.

*Kuder Preference Record.* The Kuder was given to 172 of the graduating seniors

Table 2

Correlations between Scores on the Strong Vocational Interest Blank (Form M) and the Four-Year College Average (N = 158)

Scales of the Strong	Mean*	Letter Grade Equivalent	S.D.	r
Individual Scale				
Engineers	42.6	B+	9.0	.03
Group Scales				
Group I (Human Science)	41.1	B+	15.9	.03
Group II (Technical)	46.9	A-	13.2	.07
Group V (Personnel)	36.8	B	9.3	.09
Group VIII (Office)	30.4	B-	9.6	.00
Group IX (Sales)	30.4	B-	8.5	-.06
Group X (Verbal)	33.8	B-	8.9	.13

\* Standard scores.

Table 3

Correlations between Scores on the Kuder Preference Record (Form BM) and the Four-Year College Average (N = 172)

Scales of Kuder	Mean	Percentile Equivalent†	S.D.	r
Mechanical	91.2	65	13.3	.16*
Computational	40.2	66	9.0	.16*
Scientific	76.9	80	10.6	.17*
Persuasive	64.5	35	13.5	-.10
Artistic	50.5	65	12.5	-.18*
Literary	49.8	60	14.0	.21**
Musical	18.2	61	8.6	.08
Social Service	64.8	30	16.0	-.09
Clerical	42.3	21	10.9	.04

\* Significant at the 5 per cent level.

\*\* Significant at the 1 per cent level.

† Based on male adults (1946 profile sheet).

(13 in Chemical, 50 in Civil, 79 in Electrical, and 30 in Mechanical Engineering). The two most relevant scales on the Kuder would be mechanical and scientific. The mean scores on these two scales were 91.2 and 76.9 respectively (Table 3). The equivalent percentile ratings would be the 65th and the 80th, using the male adult norms presented in the 1946 edition of the Kuder profile sheet. The correlations between the various scales of the Kuder and the four-year college average are all low (Table 3) but a few of them are significant at the five per cent level and

one at the one per cent level.<sup>6</sup> These correlations are, of course, lowered by the same factors mentioned in connection with the correlations between scores on the Strong and grades.

#### Kuder Freshman-Senior Correlations

Thirty-two students who took the Kuder during their senior year also took it during their freshman year. In comparing the mean scores (Table 4) the only difference that is statistically significant is that for the scientific scale (mean of 81.8 as freshmen; mean of 73.5 as seniors).

The correlations between the two sets of scores vary considerably (Table 4) from one scale to another (-.22 to +.66). These correlations should be thought of not only as an index of reliability but also as an index of stability of interests over a four-year period.

Finding only a limited relationship between the ratings on the interest questionnaires and academic grades is what one would expect on the basis of other studies (1, 3, 7, 9; for summaries see 11 and 12). Of course, in a counseling situation the interest questionnaires are used with the idea of obtaining information about the interest pattern, not with

<sup>6</sup> It is interesting to note that the only scale with an r significant at the one per cent level is the Literary Scale, which is the same one Yum (14) found to be significantly related to grades made by the men in his study.

Table 4

Correlations between Scores on Kuder Obtained during Freshman and Senior Years (N = 32)

Scales of Kuder	Mean Score		Percentile Equivalent*		S.D.		r
	Fresh.	Sr.	Fresh.	Sr.	Fresh.	Sr.	
Mechanical	85.8	90.3	57	64	17.8	13.3	.14
Computational	34.7	36.8	48	55	10.5	11.7	.65
Scientific	81.8	73.5	88	72	9.9	12.8	.51
Persuasive	65.3	67.6	37	41	12.9	14.1	.27
Artistic	55.0	53.8	77	75	13.8	14.2	.66
Literary	43.5	46.9	42	51	13.6	11.9	.55
Musical	19.8	22.0	70	75	8.3	8.4	.49
Social Service	67.1	62.0	35	25	18.8	18.1	.57
Clerical	42.3	37.8	21	12	13.6	12.3	-.22

\* Based on male adults (1946 profile sheet).

the idea of getting information that will help to predict academic grades.<sup>7</sup>

### Summary and Conclusion

Using a weighted grade-point average based on four years of college work as a criterion the results of this study indicate that the selection of freshman engineering students can be improved by the use of both high school averages and test scores. The effectiveness of the following tests were investigated: Scientific Verbal Ability, Comprehension of Scientific Materials, and General Mathematical Ability.

The correlations found between two interest questionnaires (Strong and Kuder) and college grades are not high enough to warrant the inclusion of ratings on such questionnaires in a selection battery, but yet it is felt that such instruments are useful in an individual counseling situation.

Received January 15, 1953.

### References

1. Berdie, R. F. The prediction of college achievement and satisfaction. *J. appl. Psychol.*, 1944, **28**, 239-245.
2. Crawford, A. B. and Burnham, P. S. *Forecasting college achievement; Part I: General considerations in the measurement of academic promise*. New Haven, Conn.: Yale Univ. Press, 1946.
3. Holcomb, G. W. and Laslett, H. R. A prognostic study of engineering aptitude. *J. appl. Psychol.*, 1932, **16**, 107-115.
4. Kandel, I. L. *Professional aptitude tests in medicine, law, and engineering*. New York: Teachers College, 1940.
5. Long, L. and Perry, J. D. Entrance examinations at the City College of New York. *Educ. psychol. Measmt.*, 1947, **7**, 765-772.
6. Lord, F., Cowles, J. T., and Cynamon, M. The Pre-Engineering Inventory as a predictor of success in engineering colleges. *J. appl. Psychol.*, 1950, **34**, 30-39.
7. Melville, S. D. and Frederiksen, N. Achievement of freshmen engineering students and the Strong Vocational Interest Blank. *J. appl. Psychol.*, 1952, **36**, 169-173.
8. Moore, J. E. A decade of attempts to predict scholastic success in engineering schools. *Occupations*, 1949, **28**, 92-96.
9. Phillips, W. S. and Osborne, R. T. A note on the relationship of the Kuder Preference Record Scales to college marks, scholastic aptitude and other variables. *Educ. psychol. Measmt.*, 1949, **9**, 331-337.
10. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford, Calif.: Stanford Univ. Press, 1943.
11. Stuit, D. B., et al. *Predicting success in professional schools*. Wash., D. C.: Amer. Council on Educ., 1949.
12. Super, D. E. *Appraising vocational fitness*. N. Y.: Harper, 1949.
13. Vaughn, K. W. The Pre-Engineering Inventory. *J. engng. Educ.*, 1944, **34**, 615-625.
14. Yum, K. S. Student preferences in divisional studies and their preferential activities. *J. Psychol.*, 1942, **13**, 193-200.

<sup>7</sup> See discussion by Strong (10, pp. 17-19).

## Study of Values Profiles Adjusted for Sex and Variability Differences

Julian C. Stanley

Department of Education, University of Wisconsin

In 1951 the 20-year-old Allport-Vernon *Study of Values*, a scale for measuring evaluative attitudes, was revised by Allport, Vernon, and Lindzey (1). Especially, its social scale was altered considerably in an attempt to secure greater homogeneity.

As originally, the average score of the standardization group on *each* of the six values has been equalized, now approximating 40 for men and women combined. Marked systematic sex differences with respect to both means and standard deviations remain, however. The women are more religious, aesthetic, and social; the men are more theoretical, political, and economic. The range of the 12 means listed in the Manual of Directions (1) is nearly seven raw-score units, while variances go from 37 to 111.

Because each testee has exactly 240 points to allot, there is no individual profile level, since the mean of his six value scores *must* be 40. The revised booklet supplies only one norm profile, employing a mean of 40 for any value for either sex.

If a profile is to be used at all, it seems

desirable to remove the group level factors—mean *and* standard deviation—separately for each sex. This has been done in Table 1 for the 1,816 college students who make up the general norms. Note there, for example, that a raw theoretical score of 53 has the same standard-score meaning for men as a theoretical score of 46 has for women. Furthermore, 43 on theoretical is for men equivalent to 37 on aesthetic.

Two cautions are appropriate here. First, Table 1 is based upon national norms and may therefore be somewhat imprecise in certain local situations. For example, when scanning profiles we should remember that both women and men in the Southeast may tend to score higher on the religious scale than do those in other sections (2,3). Second, as the *Study of Values* authors warn (1), a "high" score is high in an *inter-individual* sense only if comparisons are made among persons who can reasonably be expected to have the same average value level. Therefore, interpretations should usually be confined to the relative prominence of *intra-*

Table 1\*

Centile Sheet for College Men and Women on Allport-Vernon-Lindzey *Study of Values*  
Note: Based upon the Norms (851 Men, 965 Women) in the Manual of Directions (1).

Centile	Theoretical		Economic		Aesthetic		Social		Political		Religious	
	M	W	M	W	M	W	M	W	M	W	M	W
90	53	46	54	48	50	53	47	50	51	46	50	57
75	48	41	48	44	44	48	43	46	47	42	44	50
50	43	36	42	39	37	42	38	41	43	38	37	43
25	38	31	36	34	31	36	33	37	38	34	30	36
10	34	27	30	29	25	31	28	32	34	30	24	30

\* This is an abbreviated table. To reduce costs the original table has been deposited with the American Documentation Institute. Order Document 3960 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C., remitting \$1.25 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.25 for photocopies (6 X 8 inches) readable without optical aid.

individual values, since the *inter*-individual meaning of either raw or standardized level-free scores will not be clear when heterogeneous groups are involved.

Table 1 is merely a statistical attempt to rid the scale of certain inequalities that seem to make *intra*-individual comparisons less precise.

*Received January 26, 1953.*

### References

1. Allport, G. W., Vernon, P. E., and Lindzey, G. *Study of Values: a scale for measuring the dominant interests in personality*. Booklet and manual of directions. Boston: Houghton Mifflin, 1951.
2. Gray, Susan W. A note on the values of southern college women, white and Negro. *J. soc. Psychol.*, 1947, **25**, 239-241.
3. Stanley, J. C. and Gray, Susan W. *Sex differences and self-insight on Spranger's value types*. Mimeographed, 1951.

## A Scale for Measuring Work Attitude for the MMPI

Mary Tydlaska

*Columbia-Southern Chemical Corporation, Lake Charles, Louisiana*

and Robert Mengel

*Lake Charles, Louisiana Air Force Base*

The Minnesota Multiphasic Personality Inventory (3) is one of the most recent and among the best of personality inventories. It is designed to measure many aspects of personality by scoring various combinations of items. Although this scale has found great use in its present form in clinics and hospitals, it has not been extensively used by industry in pre-employment testing. For this latter purpose it was thought desirable to determine if there were items which could distinguish between individuals whose personality organization expresses desirable attitudes toward work and good motivation toward it and individuals whose work attitude is notoriously poor.

### Selection of Subjects

Two groups of subjects were studied. Two examples of work attitude were available. The 50 subjects from the Columbia-Southern Chemical Corporation in Lake Charles, Louisiana are current employees who were given the MMPI in a program of pre-employment testing which preceded their employment. Only those employees who had completed two or more years of satisfactory work performance were included in this study. Satisfactory work performance was based on merit ratings given semi-annually and a mean score of 3 (defined as 'satisfactory') was used as the criterion.

The 60 air force 'poor work attitude' subjects whose MMPI records were used in this study were male white air force service personnel in the 806th Supply Squadron at the Lake Charles, Louisiana Air Force Base. The category 'poor work attitude' represents 43 A. W. O. L. cases, 7 disciplinary problems, 8 individuals suspected of malingering, and 2 miscellaneous cases.

The senior writer served, during the sum-

mer of 1952, as a consultant in administering and interpreting a battery of tests designed to aid the commanding officer in working with these and similar individuals. An evaluation of 'poor work attitude' for each of these 60 cases was made on the basis of one or more interviews and test data, including a sentence completion test and the MMPI.

The groups of air base 'poor work attitude' cases and 'satisfactory work attitude' employees were matched for certain items of biographical data. These variables include intelligence (an Otis IQ for the industrial employees and an Airman's Qualifying Exam score for the air base personnel), age, education, general occupational level, and marital status. The typical subject was about 27 years of age, had average intelligence, had completed the eleventh grade of school, and was more likely to be married than single.

### Purpose

The original purpose of this study was to utilize the 60 'poor work attitude' air base personnel MMPI scores as criteria to evaluate a number of MMPI items previously selected on an a priori basis by seven individuals in the field of personnel selection and testing, as representing information indicative of an applicant's work attitude. This group of seven experts was composed of three individuals teaching courses in industrial psychology and associated with a college or university. The remaining four judges were personnel or employment managers with 15 years mean experience in personnel selection.

These items were selected in the following way. Each judge was asked to indicate on an MMPI group score sheet those statements and their deviant response which would give him insight into the general motivational pattern and work attitude of an applicant for

employment. All items which were selected by as many as three out of seven judges were included in the experimental form of the Work Attitude Scale.

This preliminary screening of MMPI items was undertaken to eliminate a number of anticipated items such as those which were found most valid in screening A. W. O. L. recidivists, "excessive use of alcohol, misbehavior in school, trouble with the law . . ." (1, p. 231). The writer was aware of the possibility that such items might be found to discriminate but postulated that they would not contribute the specific type of information which would be most valuable in helping an employment manager gain insight into a potential employee's subsequent work attitude. Most items of this nature were not selected by three or more judges and, thus, were not included in the experimental Work Attitude Scale. A total of 58 items composed the experimental scale.

A further consideration for eliminating items not selected by three or more judges was the desire to establish a number of items which would be of practical value and specific interest to employment managers in pre-employment testing. Deviant responses to these selected items can be read individually, and they can then be evaluated subjectively as well as quantitatively scored.

A technique designed to contribute more meaningful information from a normal MMPI

profile would have great utility in aiding a non-clinically oriented personnel staff member to evaluate applicants from the standpoint of their potential adjustment to an employment environment. The over-all design of this study was to provide for pre-employment testing an MMPI Work Attitude Scale tailor-made for that specific purpose.

### Procedure

An examination of the sub-scales in terms of their relationship to the two groups of individuals included an inspectional analysis of the MMPI profiles. This inspection was conducted in order to determine the number of profiles classified as normal and the number having T scores at 70 on one or more sub-scales. Table 1 presents the results of this inspectional analysis. A comparison was also made of the previously selected individual MMPI items and both groups were scored for these individual items.

### Results

Significant differences were found between the profiles of the 'poor work attitude' individuals and the 'satisfactory work attitude' employees. For example, 43 (or 71.7 per cent) of the 60 'poor work attitude' cases had one or more T scores of 70 while only 9 (or 18.3 per cent) of the 50 'satisfactory work attitude' employees had one or more scores of 70 or more.

Table 1

Inspectional Analysis of MMPI Profiles of 'Satisfactory Work Attitude' Employees and 'Poor Work Attitude' Air Base Service Personnel

	'Satisfactory Work Attitude' Employees				'Poor Work Attitude' Air Base Personnel			
			Cumulative				Cumulative	
	N	%	N	%	N	%	N	%
Normal Profile	41	82	41	82	17	28.4	17	28.4
1 sub-scale 70 or over	8	16	49	98	15	25	32	53.4
2 sub-scales 70 or over	1	2	50	100	6	10	38	63.4
3 sub-scales 70 or over					8	13.3	46	76.7
4 sub-scales 70 or over					10	16.7	56	93.3
5 sub-scales 70 or over					4	6.6	60	100
	50	100			60	100		

Table 3  
Scores on a Tentative Work Attitude Scale

Tentative Work Attitude Scale	Frequency of Satisfactory Work Attitude Employees (N = 50)	Frequency of Poor Work Attitude Air Base Personnel (N = 60)
25-29	0	8
20-24	1	15
15-19	2	18
10-14	7	12
5- 9	25	1
0- 4	14	0
Mean	7.0	16.4
S.D.	4.1	7.4

Only 37 items, from the 58 previously selected by three or more judges, were found to distinguish between 'poor work attitude' individuals and 'satisfactory work attitude' employees at the .01 level of confidence. The previously selected items in the experimental scale with the highest chi-square values were then combined and are presented in Table 2<sup>1</sup> as a Work Attitude Scale. The items are arranged in the following order: rank order in differentiating ability, the MMPI booklet number of the item, the deviant response, the MMPI item, the number of each group giving the deviant response, and the chi-square value attached to the deviant response.

The MMPI's of the two groups were re-scored in order to obtain the score each individual in the two groups made on this Work Attitude Scale. The distributions for these groups are presented in Table 3. A comparison of the scores was made. The number of responses on the Work Attitude Scale for the 'poor work attitude' cases in the validation group ranged from 5 to 29 (Mean 16.4, S.D. 7.4) while scores for the 'satisfactory work attitude' employees ranged from 3 to 20 (Mean 7.0, S.D. 4.1).

<sup>1</sup> To save printing costs, a 3-page table listing the 37 items in the Work Attitude Scale has been deposited with the ADI Auxiliary Publications Project. Order Document No. 4080 from Chief, Photoduplication Service, ADI Auxiliary Publications Project, Library of Congress, Washington 25, D. C., remitting \$1.25 for 35 mm. microfilm or \$1.25 for 6 by 8 inch photo-copies.

A cut-off score was established where the number of mis-identifications reached a minimum. Using a cut-off score of 13, 15 per cent of the 'poor work attitude' cases and 12 per cent of the 'satisfactory work attitude' employee group were incorrectly identified.

Admittedly, the items in Table 2 comprise a tentative scale which requires further validation. Until comparative studies have been carried out, the writer wishes to emphasize the experimental nature of this scale. There is a possibility that work attitude may not be a general factor but rather may be highly specific to particular work situations. Some attrition of items could then be expected in cross validation.

The writer plans to subject these items to further validation by studying the MMPI scores of a group of men whose work performance and work attitude at Columbia-Southern have been consistently merit rated as 'more than satisfactory' and a group of men terminated because they were either dissatisfied with their assigned work or working conditions. Further study with a freshman college population is also planned.

An interesting generalization, however, can be made from the writer's experience in individually re-reading and scoring each deviant response. An unusually large proportion of 'poor work attitude' individuals expressed concern over their bodily functions and believed that they were not in good health. As this was almost a chronic complaint, it suggests that a relationship exists between the Hypochondriasis Scale and the proposed Work Attitude Scale.

The writer believes, however, that further validation of this scale would prove definitely advantageous for the purpose of screening out those individuals whose Work Attitude score suggests that they are poor risks for employment. The problem of probable risk is an important one in an employment situation. Some of the resulting consequences of a poor work attitude are: (a) loss of productive time; (b) loss of time and effort expended in training a poor worker; and (c) negative influence of a low morale worker on fellow workers.

If an applicant is hired for permanent employment, it should be with the knowledge that his work attitude will enable him to contribute positively to the demands of the work situation and environmental needs of his co-workers. It may be that such a short scale could have wide use in screening applicants in the pre-employment situation.

#### Summary

From 58 MMPI items originally selected by three or more judges working in the area of personnel selection and testing as representing insight into a potential employee's inner motivation and work attitude, 37 items were found to distinguish at the .01 level of confidence between a group of 60 male white 'poor work attitude' air force personnel and a group of 50 'satisfactory work attitude' industrial employees equated in terms of education, sex, intelligence, age, occupation, and marital status.

When the 37 items which distinguish at the highest level of confidence are combined into a scale with unit weights and using 13 as a critical score, a Work Attitude Scale is obtained which correctly identified about 85 per cent of 'poor work attitude' cases and 88 per cent of 'satisfactory work attitude' employees.

Received January 26, 1953.

#### References

1. Clark, J. H. Application of the MMPI in differentiating A.W.O.L. recidivists from non-recidivists. *J. Psychol.*, 1948, 26, 229-234.
2. Gough, H. G., McClosky, H., and Meehl, P. E. A personality scale for dominance. *J. abnorm. and soc. Psychol.*, 1951, 47, 360-367.
3. Hathaway, S. R. and Meehl, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: Univ. Minn. Press, 1951.
4. Wiener, D. N. Subtle and obvious keys for the MMPI. *J. consult. Psychol.*, 1948, 12, 164-170.

## The Effects of Experience and Change of Job Interest on the Kuder Preference Record<sup>1</sup>

Frederick Herzberg

*Psychological Service of Pittsburgh*

and Diana Russell<sup>2</sup>

*University of Pittsburgh*

Valid occupational interest patterns provide one of the major bases of vocational counseling. The usefulness of such profiles depends largely upon the nature of the sample used for their construction. Two factors, success and satisfaction of employees in a field, have been shown to affect the form of such profiles.

A study by Hahn and Williams (4) with Marine Corps women revealed that certain of the Kuder scales distinguished satisfied from dissatisfied clerical workers. DiMichael and Dabelstein (2) added to the findings by recording significant relationships between the degree of satisfaction with their employment expressed by vocational rehabilitation counselors and their interests on the Kuder Preference Record. In a report by Barnette (1) occupationally successful and unsuccessful counseled veterans were distinguished on the basis of their Kuder measured interests.

Two other possible influencing factors which may alter the values of occupational interest profiles are the experience on the job and the lack of major interest in other fields of individuals in the sample on which a particular vocational pattern is based. These factors have been little considered in the development of occupational Kuder preference norms. Slight evidence exists as to the similarity of interest patterns between experienced and inexperienced workers in various occupations. If the profiles of these two groups differ, the practice of utilizing experienced persons as the basis for vocational counseling is seriously handicapped.

The desire to remain in the same occupation may appear to be similar to satisfaction; however, many people who do not express discontent for their present occupation may nevertheless show a preference for employment in some other area. How much of an effect such an expression has on occupational norms has not been determined.

It was the purpose of this study to examine the similarities and differences on Form BI of the Kuder Preference Record: (a) between individuals experienced in various occupations and persons entering these same occupations; and (b) between individuals expressing an interest in an occupational area other than that in which they are experienced and those persons in the same field who profess no other occupational interest.

### Method

Psychological Service of Pittsburgh, in different phases of its services, has accumulated scores on the industrial form of the Kuder Preference Record for various occupations. All subjects were male adults whose interests and abilities have been measured as part of a larger testing program to select persons for promotion or employment.

At the time of an initial interview, the subjects were asked to indicate the nature of the work in which they were presently engaged as well as the positions for which they were applying. The members of each occupational group were then classified as: (a) entering the vocation for the first time (entry group); (b) having had previous experience in the area in which they were seeking employment (experienced group); or (c) seeking employment in a field other than the one in which they were experienced (other interest group). Subjects representing five occupations were

<sup>1</sup> This research was supported by a grant from the Buhl Foundation.

<sup>2</sup> Miss Russell is now on the staff of the Department of Child Study and Research, School District of the City of Erie, Pa.

Table 1  
Breakdown of Sample Studied

	Occupation				
	Engineering DOT 0X74	Sales DOT 1X55	Laboratory DOT 0X70	Managerial DOT 0X84	Laboring DOT 6X669
Entry (Inexperienced)	131	36	12	—	—
Experienced	123	82	29	27	49
Experienced with other Job Interest	12 (sales)	28 (various)*	—	20 (various)**	39 (machinist)

\* Journalism 1, Business Relations 4, Engineering 5, Office Management 1, Routine Recording 1, Structural Crafts 1, Laboring Jobs 8.

\*\* Sales 9, Engineering 4, Office Management 3, Drafting 1, Mechanical Repair 1, Machinist 2.

chosen for sampling: engineers, salesmen, production managers, laboratory workers, and laborers. The samples used in this study with the various breakdowns are shown in Table 1.

Means and standard deviations of each interest scale were computed for all sub-groups of each occupational area. The significance of any differences between mean scores of the sub-groups was determined utilizing the "t" test. A difference was accepted as significant if the "t" value was beyond the 1% confidence level.

### Results

**Engineers.** Entry engineers are found to have a higher mean on the Mechanical and Scientific scales and a lower mean on the Musical scale than experienced engineers. Engineers seeking sales positions have significantly higher Persuasive interests than experienced engineers. See Table 2, A.

**Salesmen.** Entry and experienced salesmen show similar interest profiles with no significant mean differences between them. Salesmen applying for jobs in other occupational areas record a significant drop in Persuasive scores. The larger standard deviation for the "other interests" sales group on the Mechanical scale is understandable in view of the variety of other occupations included. See Table 2, B.

**Laboratory Workers.** A comparison of entry workers with experienced laboratory workers reveals a significantly higher Scien-

tific mean for the entry group. See Table 2, C.

**Production Managers.** A higher Persuasive mean is found for production managers with other occupational desires when compared with production managers seeking employment in the same area. Almost one half of the "other interest" group were sales applicants. See Table 2, D.

**Laborers.** The Mechanical and Scientific means for laborers with machinist ambitions are significantly larger than the corresponding means for laborers desiring to continue in laboring jobs. The variances between the two groups on the Mechanical scale are significantly different but it is unlikely that the "t" ratio is produced entirely by the differences in variances (3). See Table 2, E.

### Discussion

Two generalizations are suggested by the results presented. With respect to the effect of experience on Kuder occupational norms, the interest patterns of entry and experienced workers are essentially similar. The entry groups are often characterized by higher scores on those interest scales which particularly belong with their vocational fields. Thus, it was shown that entry engineers were significantly higher on the Mechanical and Scientific scales and entry laboratory workers higher on the Scientific scale than were their experienced counterparts. These higher scores for the entry group may stem from their recent completion of training and their pre-

Table 2

A Comparison of the Means and Standard Deviations between the Competitive Groups

Groups		Mech	Comp	Sci	Pers	Art	Lit	Mus	SSer	Cler
<i>A. Engineers</i>										
Entry Engineers (N = 131)	(M)	54.1	29.4	41.0	40.3	19.8	17.8	9.0	39.0	35.4
	( $\sigma$ )	9.3	7.9	7.6	12.7	7.9	8.4	5.9	10.9	10.3
		**		**				**		
Experienced Engineers (N = 123)	(M)	48.8	30.9	38.1	42.4	20.4	18.8	11.0	41.4	37.4
	( $\sigma$ )	9.2	8.2	8.6	13.9	7.9	7.4	6.6	11.6	10.4
				**					**	
Engineers with Sales Interests (N = 12)	(M)	47.7	27.3	37.8	56.6	18.6	23.9	12.8	34.8	31.2
	( $\sigma$ )	7.9	6.8	7.8	9.7	7.6	7.1	5.5	11.3	6.6
<i>B. Salesmen</i>										
Entry Salesmen (N = 36)	(M)	36.6	21.6	32.0	61.9	18.5	21.6	16.0	42.0	35.9
	( $\sigma$ )	10.6	6.6	11.5	8.6	8.3	6.9	6.0	8.9	9.8
Experienced Salesmen (N = 92)	(M)	38.4	22.7	30.2	62.0	20.8	19.0	13.5	41.9	38.0
	( $\sigma$ )	13.0	8.4	9.6	9.5	7.5	7.8	6.7	10.8	11.5
				**						
Salesmen with other Job Interests (N = 28)	(M)	36.8	23.8	31.5	53.2	20.7	20.7	12.7	43.4	38.8
	( $\sigma$ )	18.5	9.2	9.8	9.8	7.0	8.1	6.6	10.9	12.3
<i>C. Laboratory Workers</i>										
Entry Laboratory Workers (N = 12)	(M)	48.1	27.2	52.8	42.3	19.2	16.0	11.1	41.1	31.4
	( $\sigma$ )	6.7	11.8	5.2	14.1	10.2	6.6	4.9	7.0	7.6
				**						
Experienced Laboratory Workers (N = 29)	(M)	46.8	28.7	46.1	40.3	20.0	20.2	13.7	37.2	33.6
	( $\sigma$ )	8.7	8.1	7.7	15.8	7.3	9.2	5.7	11.9	9.1
<i>D. Production Managers</i>										
Production Managers (N = 27)	(M)	52.9	25.1	41.0	36.1	21.3	16.1	11.3	40.7	39.0
	( $\sigma$ )	8.9	10.8	7.2	9.9	8.1	10.1	6.8	9.6	14.3
				**						
Production Managers; other Job Interests (N = 20)	(M)	50.7	26.2	35.1	48.0	19.1	18.2	11.2	42.5	34.9
	( $\sigma$ )	8.5	7.4	8.3	12.7	7.3	7.8	6.0	10.5	9.3
<i>E. Laborers</i>										
Laborers (N = 49)	(M)	49.4	26.4	35.3	37.2	24.6	17.3	9.9	42.5	39.4
	( $\sigma$ )	11.6	8.2	8.1	11.5	10.4	8.7	7.0	9.5	10.0
		**		**						
Laborers Desiring Machinist Jobs (N = 39)	(M)	57.4	26.8	41.9	31.6	25.6	14.5	12.7	37.8	35.5
	( $\sigma$ )	6.2	5.4	9.6	9.2	8.0	6.2	6.2	9.1	9.4

\*\* The difference between adjacent means is significant at the 1% level of confidence.

occupation with the subject matter in the characteristic areas. In addition, it is highly probable that the results reflect a slanting of responses toward the desired occupational choice. Applicants for jobs are apt to alter or modify their responses to produce what they feel are the desirable interest patterns.

A change of vocational goals is reflected in the Kuder interest scores. The nature and

amount of such change depend upon the type of work desired by the individuals seeking new occupations. Engineers typically have predominant interests on the Mechanical and Scientific scales with an average Persuasive interest. For those experienced engineers desiring sales work, the dominant interest is in the Persuasive area. A similar shift toward Persuasive interest occurs for production

managers desiring other types of jobs. Almost half of the "other interest" group were sales applicants and their weight in the group accounted for the significant change in this scale. Contrariwise, experienced salesmen seeking employment outside of sales work show lower scores in Persuasive interests. This trend is also observed when comparing laborers with machinist ambitions with laborers content to fill laboring jobs. In this instance the groups are decidedly differentiated in the expected area of Mechanical interest.

The explanation of consciously biasing responses toward the desired area mentioned before would apply more specifically to the "other interest" groups. The reason for desiring to change job areas is not known. Dissatisfaction with their present vocation could possibly have been the deciding factor with many persons in the "other interest" groups. Nevertheless, the factor of other job desires does become an important variable in the selection of samples for occupational norms.

#### Summary

1. This study was designed to determine the similarity of Kuder interests between: (a) entry and experienced workers; and (b) experienced workers and experienced workers with new occupational goals.

2. The interests of the entry groups are basically similar to those of experienced persons in the same occupation. The differences found are in the direction of higher scores for the entries on scales typical of the occupa-

tional area. This similarity lends validity to the practice of using interest profiles based on experienced workers for vocational counseling.

3. It has been shown that Kuder interest scores of persons seeking employment in a new area differ from persons in similar occupations who choose to remain in their present vocational field. The particular scale in which the differences occur follow the type of work to which the change is being made. Though definite conscious slanting of test responses occurs in a situation in which employment is involved and the reason for many of the job changes may have been dissatisfaction with their present type of work, the differences found do suggest the importance of no other job interest as a criterion in the selection of samples for determining occupational interest norms.

Received January 26, 1953.

#### References

1. Barnette, W. L., Jr. Occupational aptitude patterns of selected groups of counseled veterans. *Psychol. Mon.*, 1951, 65, No. 5 (Whole No. 322).
2. DiMichael, S. G. and Dabelstein, D. H. Work satisfaction and work efficiency of vocational rehabilitation counselors. *Amer. Psychol.*, 1947, 2, 342-343. (Abstract)
3. Fisher, R. A. *Statistical methods for research workers*. (7th Ed.) London: Oliver and Boyd, 1938. Pp. 129-130.
4. Hahn, M. E. and Williams, Cornelia T. The measured interests of Marine Corps Women Reservists. *J. appl. Psychol.*, 1945, 29, 198-211.

## Effect of Viewing Angle and Parallax upon Accuracy of Reading Quantitative Scales \*

Jerome Cohen

*Antioch College*

James M. Vanderplas and William J. White

*Aero Medical Laboratory, Wright-Patterson Air Force Base, Ohio*

An important condition which affects the accuracy of reading instruments that present quantitative information is the orientation of the instrument with respect to the observer's line of sight. When an instrument is displaced laterally from the point immediately in front of the observer, the viewing angle is decreased, and if the pointer and the plane of the dial are also displaced parallax is introduced. It is well known that decreases in viewing angle and the introduction of parallax affect reading accuracy. Manufacturers of precision instrument scales take considerable pains to eliminate these factors on scales which are to be read to close tolerances. In more common situations, however, where several instruments are displayed on a flat panel, as in aircraft, it is not feasible to construct instruments with precise pointer-locating devices on them, such as mirrors, etc. The usual design practice in such situations has been to restrict the location of instruments that require great reading accuracy to the center of the instrument panel, thus avoiding the problem of parallax for at least some instruments.

Whether this latter practice is necessary or desirable has been the topic of discussion by several investigators concerned with instrument dial legibility and design practice. Recommendations by Calvert (3) and Du Bois (4) suggest that when aircraft instruments must be displaced laterally to the observer's forward line of sight, part or all of the instrument panel (or each instrument dial face) be tilted so the dial faces are perpendicular to the line of sight. That this kind of arrangement creates new difficulties

has been pointed out by Barr (1), who, while agreeing that readability could be increased by tilting the dial faces or curving the instrument panel, mentions that space requirements often make it impossible, lighting and reflection problems are more difficult, and new hazards are created (e.g., fouling during emergency escape procedure). Kappauf (6), however, suggests that for instrument panels where space limitations and other factors are not a serious consideration, panel shape and instrument dial orientation might well be a subject of careful study.

It appears to be generally agreed that excessively oblique views of instrument dial faces create serious reading errors and that such situations should be avoided in the design of instrument panels. But while these conceptions, based upon experience with dial and panel design, are sound, very little empirical evidence or theory exists to define precisely the limits of reading accuracy which might be expected as a function of viewing angle and parallax. Some data on the subject have been accumulated by Bartlett and Mackworth (2) in a study of errors made in locating aircraft position when represented on a plotting board grid. These investigators collected extensive data on the number of gross location errors made when the plotting board was seen from various viewing angles and distances. But while these data suggest that decreasing viewing angle beyond a critical point (about 35 degrees) results in gross errors, the results are not directly applicable to an instrument reading task, due to the nature of the apparatus used and the conditions of their experiment.

It would appear from the Bartlett and Mackworth data that reading accuracy can be expected to decline systematically as a

\* The experiments reported here were performed at Antioch College, Yellow Springs, Ohio under Air Force Contract No. AF 18(600)-50.



DIAL TYPE 1  
600 X 10



DIAL TYPE 2  
400 X 10

FIG. 1. Dial types used in experiment I.

function of decreasing viewing angle and that serious limitations on instrument location need to be imposed if reading accuracy is not to suffer. It appears desirable also to determine what changes in accuracy could be expected in the more special case of instrument dials, in order to discover if any generality exists for present empirical findings or for certain postulated invariants. It is the purpose of this paper to report on two exploratory studies designed to determine the changes in reading accuracy which might be expected to occur as a function of decreasing viewing angle and the introduction of parallax.

### The Experiments

The first experiment was designed to determine the effects of changes in viewing angle<sup>1</sup> upon reading errors without parallax effects entering into the situation. Photographs of dials were used to rule out the effects of parallax. The photographs were presented in a tachistoscope and read at various viewing angles from 90 degrees to 25 degrees. A total of 20 college students were used as subjects in the experiment. All had normal Snellen acuity and none had obvious visual defects.

The apparatus consisted of a sliding mirror tachistoscope with a switching mechanism so that the subject could control the exposure time. The back of the tachistoscope was arranged so that the photographs could be

tilted either horizontally or vertically and presented at all viewing angles between the limits used. Ten of the subjects read the dials tilted horizontally, and ten read them tilted vertically. Viewing distance was 28 inches, and the brightness of the white parts of the dials was seven foot-lamberts.

Two kinds of dials, as shown in Figure 1, were used in the experiment. One was a 600 X 10 dial, and the second was a 400 X 10 dial. Each subject was given ten practice trials to familiarize him with the dials and the apparatus. He was instructed to read the dials to the nearest five units as accurately and quickly as possible. Each subject was given 40 test trials, 20 on each of the two dial types, four at each of ten viewing angles used. Five of the pointer settings for each dial type were presented in each of the quadrants of the circle. For each dial, half of the settings were nearest a graduation mark, and half were nearest a mid-mark position. On each trial the subject pressed a switch, opening the shutter and exposing the dial. When he had read the setting, he released the switch, closing the shutter, and reported the apparent setting to the experimenter, who recorded the setting report and the time.

*Results of the First Experiment.* Analyses of the data revealed no systematic change in reading time as a function of viewing angle within the limits studied for either dial type or direction of slant. A fairly systematic trend exists, however, for errors with decreasing viewing angles. A graph of the per-

<sup>1</sup> By viewing angle we mean the acute angle formed by the intersection of the plane of the dial face and the observer's line of sight.

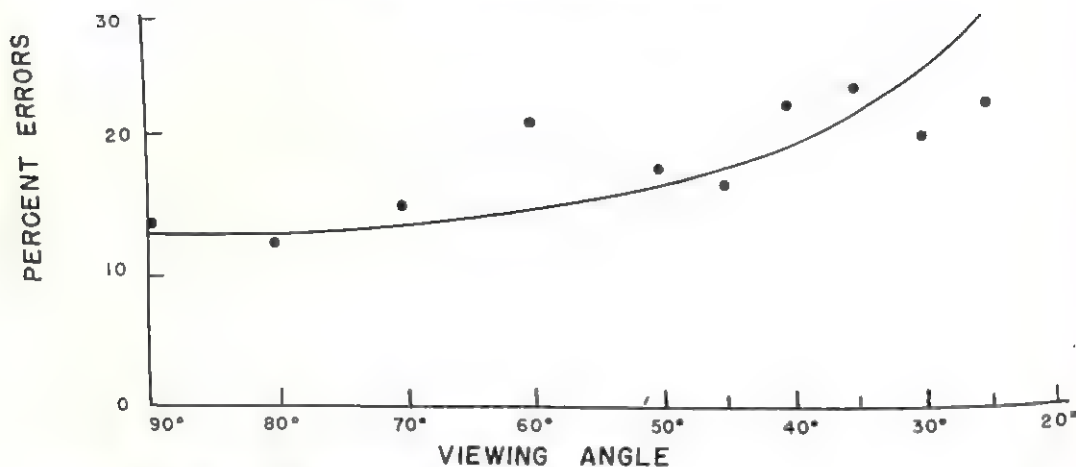


FIG. 2. Per cent readings in error by 5 units or more at each viewing angle. Experiment I.

centage of the total readings in error by five or more units made at each viewing angle by all the subjects is presented in Figure 2. The trend for errors to increase with decreasing viewing angles is represented by a cosecant function. This trend might also be represented equally well by a straight line or many other possible functions; the reason that a cosecant curve was used will be discussed later. The data for both dials and both directions of slant are combined, since there were no real differences between the two dials or directions.

Before discussing the results of the above described experiment, let us first turn to the second experiment, on the effects of parallax. The second experiment was designed to isolate, if possible, the effects of the introduction of parallax<sup>2</sup> in various amounts upon errors in pointer location. It was felt that the effects of parallax could be studied best in a situation which was fairly simple and in which errors as interpolation, numeral identification and other factors were not present. Apparatus was therefore designed so that the subjects could align a pointer with a single mark against a homogeneous background, on the assumption that the setting errors thus measured would reflect the errors in perceived location.

The apparatus for this experiment consisted of a white cardboard upon which was

<sup>2</sup> By parallax we mean the amount of displacement of the pointer from the plane of the dial face.

painted a single black line three inches high by one-eighth inch wide. A black pointer of similar dimensions was set in front of the mark on a track to permit the pointer to slide laterally to various positions in front of the mark and background. The subjects controlled the position of the pointer by pulling alternately on a pair of strings. The experimenter could read the position of the pointer from behind the apparatus by referring to a meter stick calibrated to the pointer position. A white cardboard screen allowed the subject to see only the background, black mark and pointer.

Seventeen college students acted as subjects. They were instructed to set the pointer on a line perpendicular to the plane of the mark and in line with the mark. The panel was set at various viewing angles between 90 degrees and 20 degrees inclusive, both left and right. Pointer and mark were displaced by distances of  $\frac{1}{4}$ ,  $\frac{1}{2}$ , 1, and  $1\frac{1}{2}$  inches. Viewing distance was ten feet. Two settings, left and right, were made by each subject at each of nine viewing angles, left and right, and at each of the four displacements, a total of 136 settings per subject. Sequences of presentation of the viewing angles, displacements, and left and right settings were randomized for each subject.

*Results of the Second Experiment.* A preliminary analysis of the error data revealed that the distributions of errors at each viewing angle were neither normal nor homogeneous

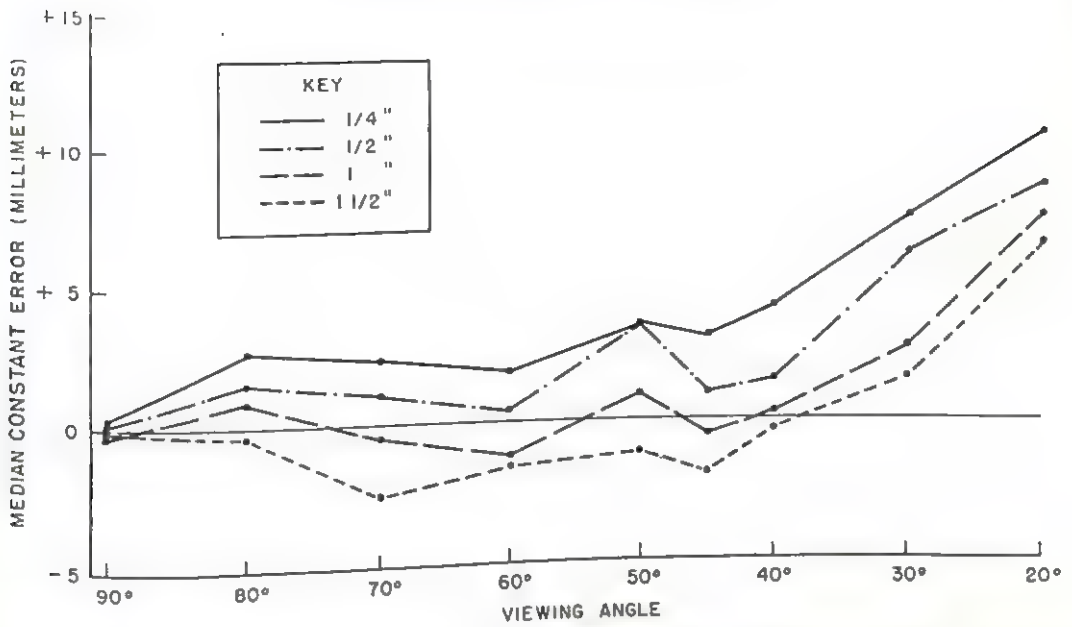


FIG. 3. Median constant error for each pointer-mark displacement at each viewing angle. Viewing distance was 10 feet.

with respect to variability. The distributions were symmetrical at the 90 degree position, but as viewing angle was decreased the distributions became both more skewed and more variable. For this reason the medians were used rather than the means as average

measures of errors at each viewing angle and displacement.

Figure 3 represents the data of the experiment, and is plotted in terms of the median constant error (in which the direction of error is considered) as a function of viewing

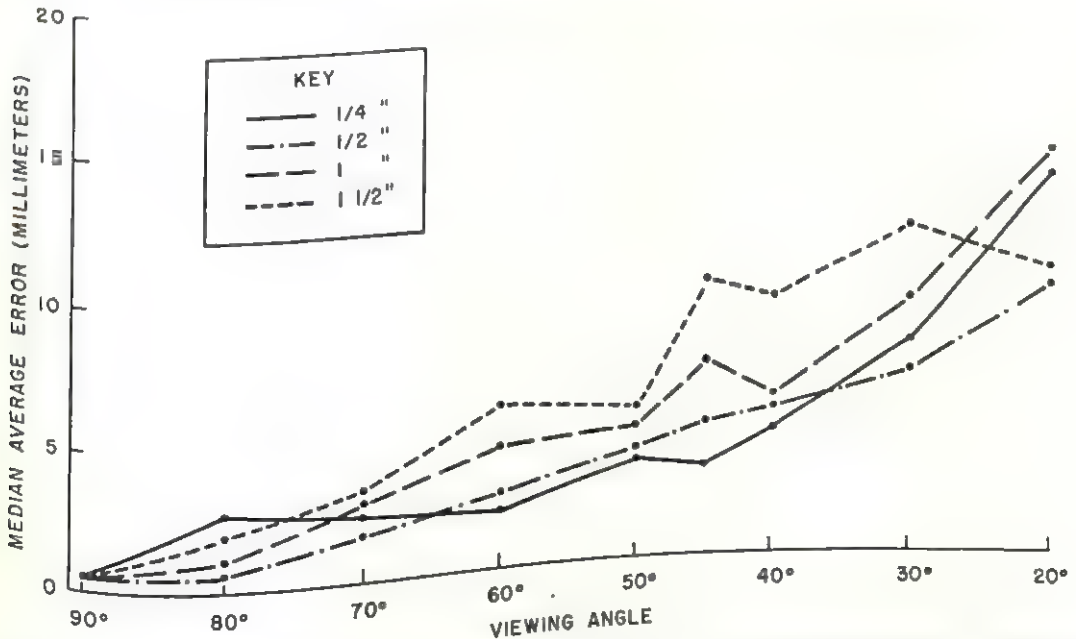


FIG. 4. Median average error for each pointer-mark displacement at each viewing angle. Viewing distance was 10 feet.

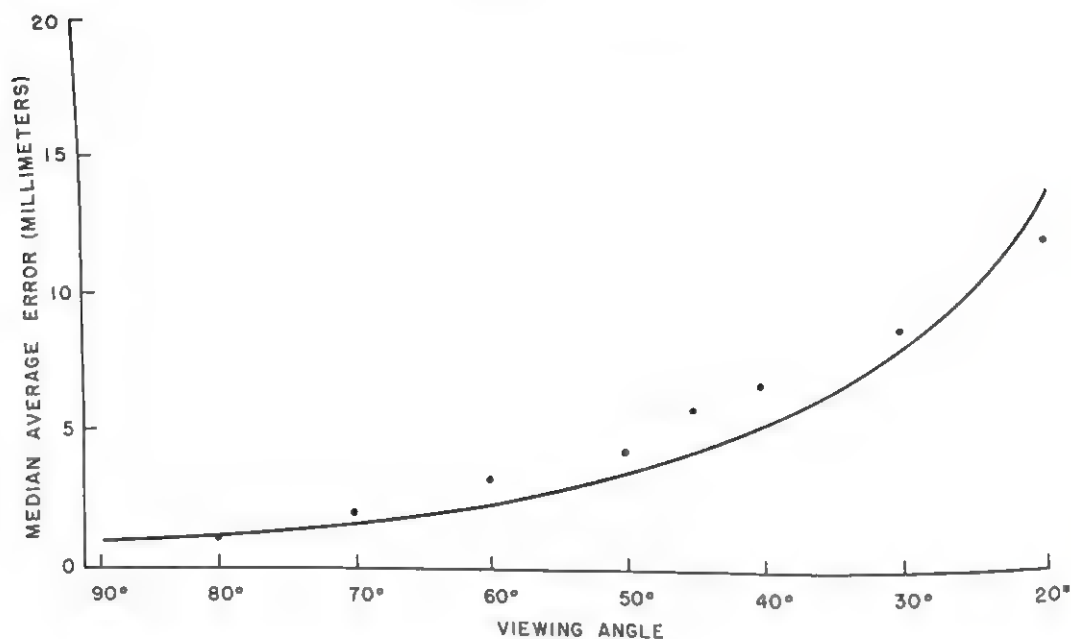


FIG. 5. Median average error for all pointer-mark displacements at each viewing angle. Viewing distance was 10 feet.

angle, for the four pointer-mark displacements used. The zero point indicates correct alignment, positive values indicate settings in the direction away from the observer, and negative values indicate settings toward the observer, relative to correct alignment.

It can be seen from Figure 3 that a trend exists here similar to that found in the first experiment, namely, increasing error with decreasing viewing angle. In addition, there appears to be an inverse relation between constant errors and displacement, the small displacements yielding apparently greater positive constant errors, while the large displacements yield small or negative constant errors (in the direction of the observer).

A similar trend appears for errors to increase as a function of viewing angle if the median average error (in which direction is not considered) is plotted. Figure 4 contains a graph of these data plotted for each of the four displacements. The relation of amount of displacement to error is not so clear here as it was for constant error. The amount of the error as indicated by the median constant error, appeared to be inversely related to the amount of displacement. In Figure 4, it appears that the greater the

displacement, the greater is the error variability, as measured by the median average error.

If all the data are combined and the median average error is plotted as a function of viewing angle for all displacements combined, the data as shown in Figure 5 appear. The curve fitted to these data is again a very close approximation to a cosecant function; it appears to fit the data quite well and in about the same way as in the first experiment.

### Discussion

There appear to be at least two effects involved in the experimental data. The first may be termed the effect of decreased apparent distance between successive points on the scale as a function of decreasing viewing angle; the second appears to be the effect of displacement between the pointer and scale plane. Apparently, for the limits of this study, large pointer displacements lead to small constant errors but large variable errors, while small displacements lead to large and consistent constant errors.

These results may be interpreted, at least in a preliminary way, by considering the visual angle relationships which vary con-

comitantly with changes in viewing angle. A decrease in viewing angle results in a corresponding decrease in the visual angle subtended at the eye by a given mark separation. It was noted earlier that in Figure 2 the data were represented by a cosecant function. They might have been fitted by a straight line. However, since the projection of the mark separation distance decreases proportionally to the sine of the viewing angle, we would expect errors to be inversely related to the sine, or directly proportional to the cosecant, of the viewing angle. At zero degrees viewing angle, of course, the cosecant function is infinite, and we would also expect errors to be extremely large or erratic, since the mark separation, as projected on the retina, would be zero, and the dial face would probably not be visible. The interocular distance is purposely ignored in the geometrical analysis.

It would be expected from this interpretation that errors, in reading to a given criterion of accuracy, would remain substantially negligible until the visual angle subtended by the criterion distance approached or diminished below some minimum discriminable angle. To compensate for this decreasing visual angle, it may be necessary, in a dial reading situation, only to increase the mark separation so that the visual angle subtended by the distance representing criterion error tolerance remains above the minimum discriminable.

As an example of this kind of interpretation, we may consider a situation in which a normal observer is required to read quantitative scales. If the illumination is good, and if it is assumed that the accuracy of the readings does not require discriminations finer than about one minute of visual angle, the observer could discriminate points as separate if they are 0.008 inch apart at a viewing distance of 28 inches. Thus, under the conditions of the first experiment, since the mark separations were about 0.125 inch, the minimum discrimination necessary (for accuracy to the nearest five units) was about twice that of which the normal observer is capable. Rationally, gross increases in error frequency under these conditions would not be expected until the viewing angle was de-

creased to a value of about 30 degrees. This expectation is borne out reasonably well by the data. Presumably, under levels of illumination which are below that required for a discrimination of one minute of visual angle, a correction would be necessary for the decrease in acuity at the lower level.

It should be clear, of course, that the above interpretation may be applied only when parallax is not present (i.e., when scale marks and pointer are not displaced). It was seen from the results of the second experiment that different amounts of parallax lead not only to different constant errors (both in direction and amount), but to different variability as well. It can be seen from an examination of Figure 3 that there was a consistent tendency for the subjects to set the pointer too far away, to "overcompensate," for the amount of parallax present and that this tendency became more pronounced with decreasing viewing angle. Tentatively, a definition has been constructed for parallax in terms of the visual angle subtended at the eye by the pointer-mark displacement distance, and a theoretical function has been derived to account for these kinds of errors. A study is planned to test this theoretical function in a scale reading experiment. It is hoped that such an approach will serve a two-fold purpose of providing a basis upon which to predict the errors to be expected in a practical sense, and at the same time provide a better understanding of the functional relationships involved.

### Summary

Two experiments were performed to evaluate and quantify the effects of decreased viewing angle and parallax upon accuracy of reading instrument scales. In the first experiment, viewing angles were varied, and subject-controlled tachistoscopic presentation of the stimulus dials was used. Dial photographs were used as stimulus materials to isolate effects of viewing angle from those of parallax. The results show that reading errors increased as viewing angle decreased from 90 degrees to 25 degrees. Reading time was unaffected by changes in viewing angle. In the second experiment the effect of

parallax was studied by requiring the subjects to align a movable pointer with a mark. The apparatus was set at viewing angles between 90 degrees and 20 degrees, and four pointer-mark displacements were used. The average error of the settings increased as the viewing angles decreased. The increase is approximated by a function proportional to the cosecant of the viewing angle. The constant error tends to increase systematically with viewing angle. With increasing pointer-mark displacement the average error tends to increase, but the constant error is inversely related to the amount of pointer-mark displacement.

The error curves in both experiments are approximated by a function proportional to the cosecant of the viewing angle. An interpretation in terms of a least discriminable visual angle is advanced to account for the results.

Received January 26, 1953.

## References

1. Barr, M. L. Visibility of cockpit instruments. *J. Aviation Med.*, 1950, 21, 328-342.
2. Bartlett, F. C., and Mackworth, N. H. Planned seeing; some psychological experiments. I. Visibility in the control of fighter command. II. The synthetic training of pathfinder air bombers in visual centering on target indicators. *Air Ministry Publication No. 3139 b*. London: His Majesty's Stationery Office, 1950.
3. Calvert, E. S. The scientific basis for the new British system of cockpit lighting. *Elect. Engng.*, N. Y., 1944, 63, 869-870.
4. Du Bois, E. F. The anatomy and physiology of the airplane cockpit. *Aeronaut. Engng. Rev.*, 1944, 4, 15-21.
5. G. B. Royal Aircraft Establishment. Layout of cockpits with particular reference to night visibility and lighting. *G. B. Royal Aircraft Establishment Electrical Engineering Dept., Memo No. 727*, February, 1942.
6. Kappauf, W. C. Studies pertaining to the design of visual displays for aircraft instruments, computers, maps, charts, tables, and graphs: a review of the literature. *USAF, Air Materiel Command AFTR No. 5765*, April, 1949.

# Visual Tracking: III. The Instrumental Dimension of Motion in Relation to Tracking Accuracy<sup>1</sup>

Robert S. Lincoln

*The Johns Hopkins University*

In the typical tracking task a human operator is required to make corrective responses to visual cues provided by the relative positions and rates of travel of a target and a cursor or follower. With modern equipment the operator seldom manipulates the cursor itself. Rather he effects his control through a mechanical or electrical device that links him to the controlled cursor. This connecting mechanism usually alters the relationship between the operator's controlling motions and the actual movements of the cursor. The nature of the alteration is determined by the characteristics of the tracking instrument. Systematic variation of these characteristics defines the instrumental dimension of tracking motions.

Within this dimension three different types of instrumental alteration of motion are of special interest since all of them have been incorporated in remote-control tracking equipment. These alterations may be termed: (a) translations, (b) transformations, and (c) integrations of motion. Translated motions directly reflect the characteristics of the motions made by the operator, but the translated motions are either amplifications or reductions of the operator's motions. Transformed motions reflect the operator's movements only in a special way since the system output is not a direct counterpart of the motion input (2). In the tracking device considered in this paper, the transforming instrument changes an input of extent of movement into an output of velocity of movement.

Integrations of motion involve the combination into one output of a simultaneous translation and transformation of the same movement of the operator.

The various instrumental alterations of motion that have been described are produced by direct, velocity, and aided tracking systems, respectively. Table 1 indicates the component motions required to achieve control of the position and rate of travel of the cursor in each of the three types of tracking, and the alterations of these motions that are produced by the different tracking systems.

As Table 1 shows, direct tracking involves two translations of motion while the velocity tracking mechanism produces two transformations of motion. In aided tracking an integration of the simultaneous translation and transformation of the same positioning motion is developed by the tracking device.

This study was designed to provide information related to three main questions:

Table 1  
Instrumental Alterations of the Operator Motions  
Required to Achieve Control of the Position  
and Rate of Travel of the Cursor

Type of Tracking	Positioning Motion	Rate Motion
Direct	Translated into cursor positioning	Translated into rate of cursor travel
Velocity	Transformed into direction of cursor travel	—
	Transformed into rate of cursor travel	—
Aided	Translated into cursor positioning	—
	Transformed into rate of cursor travel—the translation and transformation are integrated.	—

<sup>1</sup> This report is based on a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Wisconsin in 1952. The research was conducted under the direction of Dr. Karl U. Smith to whom the writer is greatly indebted. Support for the research was provided by the Research Committee of the Graduate School from special funds voted by the Legislature of the State of Wisconsin. Reported at the 1953 meeting of the Eastern Psychological Association.

1. Are the characteristics of the curve of skill acquisition in tracking changed by the different instrumental alterations of motion?

2. Are the effects of practice sufficient to overcome differences that may exist in the difficulty of operation of control systems that produce translations, transformations, or integrations of control movements?

3. What transfer effects appear when training on one type of tracking is followed by transfer to another type?

### Apparatus and Procedure

The apparatus used in this study has previously been described in some detail (3, 4, 5). The operator's task is to align a cursor and a moving target by means of a handwheel control. The target moves over a circular course that includes numerous reversals in the direction of target movement and continuous changes in target velocity.

In order to compare the levels of accuracy that are obtained with direct, velocity, and aided controls, a special device has been constructed that permits a rapid change from one type of tracking to another without the alteration of other critical features of the tracking task.

Tracking-accuracy scores are obtained with a mechanism which integrates the tracking error record and provides a summated accuracy score on the dial of an electric clock.

Prior to the experimental comparison of the different types of tracking, data were obtained on the optimum ratios of handwheel-to-cursor displacement (4). These optimal ratios were used in the comparison of the three types of tracking in order to insure that obtained differences in accuracy levels would not be a reflection of arbitrarily chosen displacement ratios. For all ratios used in aided tracking, an aided tracking time constant of 0.5 second was maintained (6).

Three groups of 18 subjects each were used in the study of training and transfer of training effects. Subjects were randomly assigned to training groups, and a different group of subjects was trained on each of the three types of tracking. The training sessions extended over a period of six successive days. On each training day all subjects received ten one-minute tracking trials. A 25-second rest pause was permitted between trials.

Before the training trials were begun, all subjects were given an explanation of the tracking task and of the mechanical features of the control that they would use. A brief demonstration of the control was also provided. This same explanation and demonstration was given to subjects who transferred to a new type of tracking during the transfer trials. Subjects received no information concerning their accuracy in track-

ing other than that provided by the visual display of the apparatus.

The effects of transfer were observed on the seventh day of the experiment. At this time six subjects from each of the training groups received ten trials on one of the two types of tracking on which they had not been trained. Six different subjects from each of the training groups received ten trials on the second of the two types on which the groups had received no training. The remaining six subjects in each training group continued with the type of tracking on which they had been trained. These control subjects made up the additional-practice groups. A matching procedure was used to equate the various transfer and control groups on the basis of their accuracy scores achieved during the training period.

### Results

*Results of Practice.* Performance curves for all training groups are shown in Figure 1. In this figure mean accuracy scores are plotted for each trial. The training curves indicate that those subjects who trained on direct tracking made the highest mean accuracy score on every trial throughout the entire training period. Those subjects who trained on velocity tracking made the lowest accuracy scores. The accuracy of aided tracking consistently fell between these two extremes although, after the first few trials, aided accuracy closely approached direct accuracy.

From these results it is apparent that the mechanical devices developed as an aid to the tracking operator are of no general value under the present experimental conditions. It is quite possible, however, that different results might be obtained with a more uniform target course, or in situations that require the operator to track continuously for long periods of time.

In order to evaluate the significance of some of the characteristics of the practice curves, a test of trend (1) was applied to the training data. In this test the mean accuracy scores for days, rather than trials, were used. Before the test of trend was carried out, it was necessary to perform an arc sin transformation (7) of the mean scores after each mean score had been calculated as a percentage of the maximum possible accuracy score. This procedure was required to reduce a negative correlation between the means and variances of the training groups.

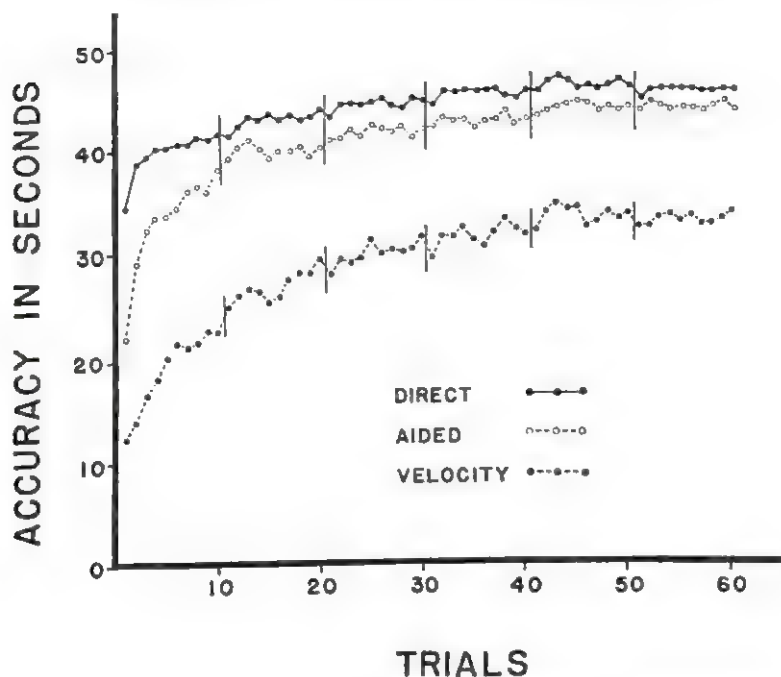


FIG. 1. The levels of accuracy reached by groups of subjects who trained on direct, aided, or velocity tracking. The vertical lines indicate the blocks of ten trials which made up a day's run.

According to the trend test, the training curves show significant ( $p < .001$ ) deviation from linearity. The curves for the separate groups also differ significantly in regard to the degree with which they depart from linearity. In addition, the slopes of best-fitting straight lines differ between groups.

*Results of Transfer.* The data obtained during the transfer trials were subjected to an analysis of variance. Before the analysis was begun, the accuracy scores were transformed in the same manner as the training scores. The analysis of variance showed all of the main sources of variation to be significant ( $p < .001$ ), but a significant interaction ( $p < .001$ ) between training and transfer types of tracking greatly modified the importance of the main variables. This result indicates that the effects of training are highly specific in nature since no type of training led to superior over-all performance when transfer was made to all three types of control. Rather the effects of training depended upon the type of tracking to which transfer was made.

As has been pointed out, direct tracking involves two instrumental translations of the operator's control motions, while the velocity

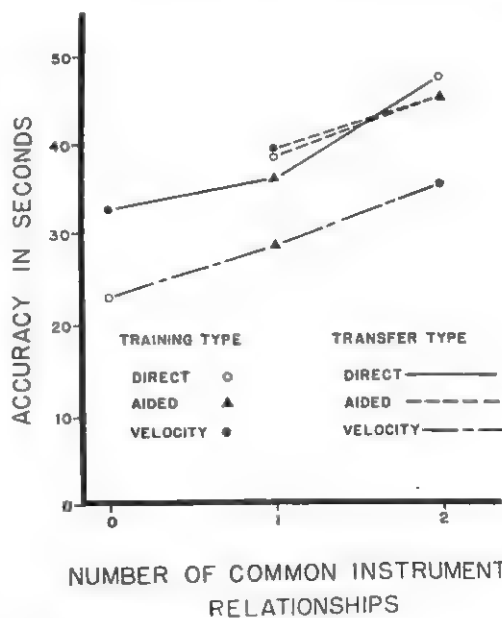


FIG. 2. Accuracy in tracking as a function of the number of common instrumental relationships between training and transfer tasks.

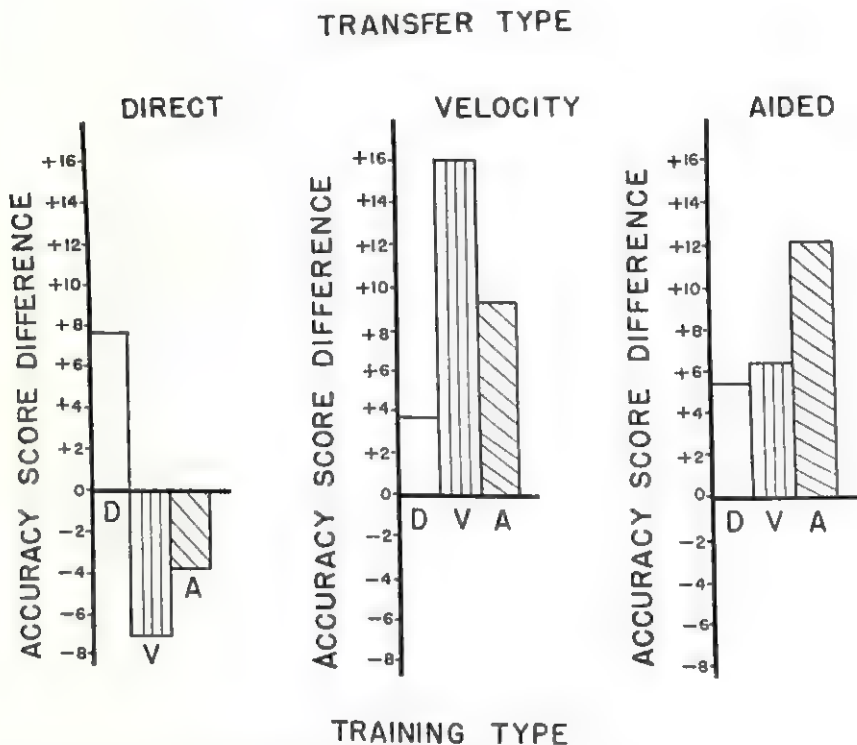


FIG. 3. Positive and negative transfer effects. The zero point on each graph represents the initial accuracy level achieved by untrained subjects on the transfer types. Plus deviations from the zero line indicate positive transfer effects, while minus deviations indicate negative transfer effects. All graphs show the type and amount of transfer effect produced when subjects are trained on direct tracking (D), velocity tracking (V), or aided tracking (A), and later transfer to the same or a different type of tracking.

mechanism produces two transformations of the operator's motions. Aided tracking involves one translation and one transformation of motion. Direct and velocity tracking, therefore, have no common instrumental relationships while aided tracking has one relationship in common with both direct and velocity tracking. Figure 2 shows that the amount of transfer, as measured by accuracy scores, is directly related to the number of instrumental relationships that are common to both the training and transfer tasks. The figure indicates (for example) that training on aided tracking led to greater accuracy upon transfer to the velocity control than did training on direct tracking. A similar interpretation may be applied to the other points on the curves.

In Figure 2 the additional-practice groups are shown as having "transferred" to the same type of control as the one on which they had

trained. The relative positions of these groups indicate that direct tracking was still slightly superior to aided tracking on the seventh day of the experiment.

Another suggestion of transfer effects may be obtained from a comparison of the accuracy scores achieved by untrained subjects on a given type of tracking with the scores made by subjects who transfer to that type following training on another type. Figure 3 pictures this kind of transfer effect. The zero point on the ordinate of each graph represents the mean score made on the first ten training trials by the 18 subjects who trained on each of the transfer types. The remaining ordinate values indicate the amount and direction of the differences between the mean scores for the untrained subjects and mean scores made by the six subjects who transferred to the different transfer types after training on either direct, velocity, or aided

tracking. In Figure 3 the additional-practice groups are again shown as having "transferred" to the type of control on which they were trained.

The significance of these transfer effects was evaluated by means of *t* tests that were performed on untransformed accuracy scores after *F* tests had indicated the homogeneity of the variances of the different training and transfer groups. All transfer effects are significant ( $p < .05$ ) with the exception of that effect produced by transfer to velocity tracking after training on direct tracking.

These results indicate that the prediction of transfer effects must take into account the direction of transfer as well as the number of instrumental relationships common to both the training and transfer tasks. Training on direct tracking produced a positive effect upon transfer to the aided control and no significant effect upon transfer to the velocity control, while training on either velocity or aided tracking produced a negative effect upon transfer to the direct control.

### Summary and Conclusions

This study was designed to provide information concerning the acquisition and transfer of skill in the operation of remote control devices which produce instrumental translations, transformations, and integrations of the operator's controlling motions. These instrumental alterations of response are produced by direct, velocity, and aided tracking systems.

Each of three groups of 18 subjects received training on either direct, velocity, or aided tracking for a period extending through six successive days. On the seventh day of the experiment 12 subjects from each training group transferred to different types of tracking while the remaining six subjects in each group continued to track with the control on which they had been trained. Accuracy scores achieved by the subjects were analyzed with regard to the effects of both practice and transfer. The results of the experiment suggest a number of conclusions.

1. The instrumental characteristics of control devices are prime determinants of the

accuracy with which those devices may be operated.

2. Practice curves for the three types of tracking show the typical negative acceleration which has previously been demonstrated in studies of direct tracking behavior.

3. For complicated target courses, the accuracy of direct tracking is consistently superior to both aided tracking and velocity tracking. Aided control is also far superior to velocity control.

4. Within the limits of this experiment, the effects of practice are not sufficient to eliminate the differences in accuracy achieved with the three types of tracking.

5. The effects of training are highly specific in nature. The best performance in transfer to any type of tracking is achieved by subjects who are trained on that specific type.

6. Negative transfer effects appear when subjects transfer from aided or velocity tracking to the direct control, while positive transfer effects appear in the reverse situation. The amount of transfer is directly related to the number of instrumental relationships that are common to both the training and transfer tasks.

Received January 12, 1953.

### References

1. Alexander, H. W. A general test for trend. *Psychol. Bull.*, 1946, 43, 533-557.
2. Craig, D. R. and Ellson, D. G. The design of controls in *Human Factors in Undersea Warfare*. Washington, D. C.: National Research Council, 1949.
3. Lincoln, R. S. and Smith, K. U. Transfer of training in tracking performance at different target speeds. *J. appl. Psychol.*, 1951, 35, 358-362.
4. Lincoln, R. S. and Smith, K. U. Systematic analysis of factors determining accuracy in visual tracking. *Science*, 1952, 116, 183-187.
5. Lincoln, R. S. and Smith, K. U. Visual tracking: II. Effects of brightness and width of target. *J. appl. Psychol.*, 1952, 36, 417-421.
6. Mechler, E. A., Russell, J. B., and Preston, M. G. The basis for the optimum aided-tracking time constant. *J. Franklin Inst.*, 1949, 248, 327-334.
7. Snedecor, G. W. *Statistical methods*. Ames, Iowa: Iowa State College Press, 1946.

## Identification of Cola Beverages Overseas \*

E. Terry Prothro

*Brooklyn College*

Bottling and sale of cola beverages is now taking place in many countries around the world, and consumption of these beverages has become a part of the life of inhabitants of every continent. Both Coca-Cola and Pepsi-Cola are widely sold in Lebanon, for example, and their popularity has stimulated the production of several imitations. It therefore seems worthwhile to determine whether or not consumers can differentiate the American beverages from each other and from the local colas on a basis of taste.

A series of investigations by Pronko and others indicated that subjects could not identify American colas better than chance when many different brands were presented (1), but that they could identify Coca-Cola significantly more often than chance when only the three leading brands were used (2). It was therefore decided to use only three beverages, including Coca-Cola, in this investigation, so that there would be maximum opportunity to reveal taste differences between the beverages.

### Procedure

The three leading cola drinks of Lebanon were used in this study. These three, in order of popularity are Coca-Cola, Pepsi-Cola, and Williams Champagne Cola. The American colas are bottled in local plants but according to a special and presumably secret process dictated by the parent corporations. Champagne Cola is produced by the Williams plant, a Lebanese organization which also produces many other soft drinks. This cola resembles the American drinks in appearance. It was introduced after the early success of Coca-Cola. Coca-Cola has been distributed there for nearly three years, Pepsi-Cola for six months, and Williams Cola for more than one year.

\* This study was conducted overseas while the author was teaching at the American University of Beirut.

A total of 60 students of the American University of Beirut volunteered to serve as subjects. Each subject stated that he was familiar with the taste of the beverages used, that he was not suffering from a cold, that he had no political or religious objection to any of the beverages. He was then told that he would be given each of the three colas in series, and that he was to identify each after its presentation. Approximately 2 oz. of refrigerated cola were used at each presentation. The beverages were presented in identical 6 oz. glasses. Subjects were blindfolded during the trials. Approximately one minute elapsed between each trial, during which time the subjects were asked to rinse their mouths with water. There are six possible arrangements when three colas are presented in series. Each of the six arrangements was used for ten subjects.

### Results

Although the subjects were informed that each of the three colas would be presented, some of them felt that instructions in a psychological experiment cannot be relied upon. One subject believed that a single cola was presented three times, and some believed that other colas were being presented. From Table 1 it can be seen that the most recently introduced cola (Pepsi-Cola) was named most often. This fact may be a result of the ex-

Table 1  
Identification Responses of Subjects when Presented with Three Cola Beverages

Cola Presented	Response				Total
	Coca Cola	Pepsi-Cola	Champagne Cola	Other	
Coca Cola	24	30	5	1	60
Pepsi	26	28	5	1	60
Champagne	1	6	51	2	60
Total	51	64	61	4	180

tensive advertising campaign that accompanied its introduction.

Our subjects were not able to differentiate between the two American colas. Indeed Coca-Cola was called Pepsi-Cola more often than it was named correctly. On the other hand, the subjects did identify the local cola quite well, and showed little tendency to confuse it with the American colas. If we employ the chi-squared test of significance, we find that the American colas are identified correctly only slightly, and insignificantly, more often than chance. The correct identification of Williams Champagne Cola was not attributable to chance. The superiority to chance was significant at the .001 level.

#### Summary

A total of 60 students in American University of Beirut were asked to identify Coca-

Cola, Pepsi-Cola, and a popular local cola without reliance on visual stimuli. Although these colas are widely distributed locally, and the subjects stated that they were familiar with them, it was found that the American colas could not be differentiated from each other. The local product could however be distinguished from the American brands in spite of the fact that it is an imitation of them.

*Received March 3, 1953.*

#### References

1. Pronko, N. H. and Bowles, J. W. Identification of cola beverages. III. A final study. *J. appl. Psychol.*, 1949, 33, 605-608.
2. Pronko, N. H. and Herman, D. T. Identification of cola beverages. IV. Postscript. *J. appl. Psychol.*, 1950, 34, 68-69.

## Applied Psychology in Action

### The Non-Directive Approach in Advertising Appeals

Howard D. Hadley

*Daniel Starch and Staff, Mamaroneck, N. Y.*

In recent years a new basic approach to psychotherapy has been developed. It is called the non-directive method. Because it has some implications for advertising, there may be value in describing this technique along with its counterpart—the directive method.

Non-directive psychotherapy is built around these central concepts:

1. That all individuals have the basic capacity to understand the forces in their lives which cause them unhappiness and pain. Moreover, they also can understand those forces which lead to pleasure and well-being.

2. That all individuals, by personal effort, ultimately can overcome the bad or enhance the good forces in their lives.

3. That this process is made easier, quicker, or more effective in an atmosphere that is friendly, sincere, and understanding.

In non-directive therapy, the therapist does not assume any of the usual responsibilities such as prescribing treatment or even directly defining the cure. Instead, the therapist attempts to set up with the individual a relationship, an atmosphere, in which the person may talk or act without danger of being criticized. It is one of complete acceptance. Within this tender environment, the person himself comes to understand and to re-evaluate the forces operating to make him happy or unhappy.

Directive therapy differs in that it usually requires a direct assault upon the individual's maladjustments: the person is tested, analyzed, and then told what is wrong and given a prescription for a cure. It is similar to going to a doctor only to find out you have appendicitis. He puts you in a hospital, removes the diseased organ, and your body then is able to complete the recovery. When dealing with the physical body, the doctor's obligation is often greater than the patient's.

When dealing with a person's mind, this method sometimes falls down because the person is reluctant to believe or is unable to accept the diagnosis of the therapist. The therapist can only discover and point out the path, he cannot walk it. Many persons benefit from this advice but many others do not.

In advertising it is very similar: there are two ways in which to advertise. The advertiser can *tell* the person to buy the product because it will do this or that for him. This is comparable to the directive method where the individual is told what his troubles are and how to cure them. Or the advertiser can create a friendly, sincere, and understanding atmosphere which shows the benefits of the product without direct intention to sell. This places the person in a situation where he is able to accept new ideas without threat to his old ideas.

To complete the comparison, directive therapy is similar to the direct appeal in advertising. In each case, the person to be influenced is directly told what is wrong and how to correct it. The non-directive therapy can be compared with the inferred appeal in advertising.

The inferred technique usually utilizes association of the product with very acceptable things, persons, or events. The acceptability of the associated "thing" creates an attitude in which the individual feels free to accept new ideas. Also, the acceptability of the "thing" is transferred to the product. Examples of advertisers using the inferred appeal are Modess, John Hancock, Breck, Old Gold, and Seagram's 7 Crown. Each of these advertisers avoids a direct assault upon the consumer's credibility but disarms him and then introduces a strong simply-worded message. Here are some examples from current advertising which help to illustrate the comparison. Take the Philip Morris theme, "Some-

thing Wonderful Happens—." Here is an attitude that some persons actually attain after changing to Philip Morris. At the present time, this very personal theme is put across in a very direct manner. Readers are now being *told* that "Something Wonderful Happens when they change to Philip Morris." It is entirely possible that the theme—which is excellent—might be more effective if readers were to arrive at this realization without such obvious aid. To be told about this anticipated change in the personality tends to act as a threat, which leads to resistance and withdrawal. The point could be put across showing this "Something Wonderful" happening to others and then associating it indirectly with Philip Morris.

Basically, persons earnestly want to believe advertising, but they are afraid to. This fear is the result of the possible "frustration" they may suffer because of false and misleading advertising when the product fails to live up to a person's expectations. If this happens with advertising which uses the direct appeal, the consumer blames the advertiser. If this failure occurs with a product which has used inferred appeals, the consumer is less apt to complain 1) because of lack of specific promises in the advertising and 2) because he was the one who decided what the product would do—and not the advertiser.

To remain in the cigarette field, let's take the king size cigarettes. A person who uses the regular length cigarettes is told that he is not being economical and is also remiss in attention to his health. While both of these points have the popular vote behind them, a person will tend to treat the advertising as a threat to his judgment. Such advertising is a negative approach to a negative appeal. The inferred (non-directive) appeal would show a person enjoying increased smoking pleasure from a longer smoke.

Throughout this whole comparison, there are two distinct philosophies of advertising. One of them is to "tell them" and the other is to "have them tell themselves." From evidence in psychotherapy, and from evidence presented in the October 1952 issue of this magazine [*Advertising Agency*], it appears that when credibility is lacking on the part of

the person, the latter (inferred) method is more effective.

Let's look at the record of some beer advertisers who use, in greater or lesser degrees, the inferred technique. Beer is taken as an example because beer is almost universally used, more money is spent for it each year than for milk or cigarettes. Also, the response of consumers to beer advertising is quicker and greater than for many other mass distributed and used products. From 1949 through 1951 there was a four per cent decrease in the amount of beer consumed. However, the leader, Schlitz, gained about 20 per cent. Rheingold had close to a 50 per cent gain, and Miller's almost doubled. While there may have been other factors operating, it's hard to deny credit to the type of advertising these beer companies have been using. As a matter of fact, if you look at the top brand in any product group, you will usually find that they have used the non-directive or inferred approach.

If such an approach is so good, why don't more persons use it? Here are some possible reasons:

1. Most advertisers (not agency personnel) cannot resist the temptation to "tell them"—to sell the product as an extroverted salesman would. This is a very easy pitfall into which to fall, and a hard one to leave.

2. Not everyone is able to use the inferred (non-directive) technique. To a large extent it is dependent upon the personality of the creative persons. Some persons think and act in a non-directive manner. They are friendly, sincere, and understanding. Because of these personality characteristics, they are able to create advertising which is in keeping with their temperament. Other persons are of a different turn. They are better at employing stronger, more obvious and promotional methods to get across a point. This is not intended as a criticism since there is certainly room in the advertising field for both. However, don't expect one type of person to turn out a different type of advertising. It is hard (and uneconomical) to "live a lie" and the consuming public is quick to catch insincerity in advertising.

3. The direct method still sells goods. This is a potent argument. While there are many persons who are influenced by it, it seems as though an advertiser should use the inferred approach if he wishes to be really big. For this approach appears to be effective with the greatest number of people.

To summarize the high points:

1. There is a close parallel between concepts used in psychotherapy and advertising.
2. The non-directive technique is quite com-

parable with the inferred technique in advertising as exemplified by Modess, Breck, John Hancock, Old Golds, and Seagram's 7 Crown.

3. Both techniques appear to be successful because the "patient" or "consumer" arrives at judgments by himself without being directly told.

4. Successful use of both methods is very dependent upon the personality of the persons involved—therapists or creative persons.

*Reprinted from Advertising Agency, April, 1953.*

### Reading: Stop Wasting Your Time

Take a look at your mail. In addition to the usual flood of letters, ads, memos, it probably contains a couple of newspapers, a rash of magazines, an occasional book or two, and a shower of releases, pamphlets, broadsides, etc. Management personnel spend about 15 hours a week just reading. And in many jobs you may well spend more. But how much of it is wasted? Tests show the average businessman reads only slightly better than an eighth-grade schoolboy—and that is still above the national level. Trouble is, few people have ever received any reading training after elementary school. More and more executives are aware of this handicap, are turning to reading development firms like The Reading Laboratory in New York and Chicago's Foundation for Better Reading. These groups specialize in training executives to read faster, better. Goal of the 20-hour course is a reading speed of 650–700 words per minute. National average is about 250 words. Many taking the course do far better.

Reading Laboratory Director K. P. Baldridge points out that reading speed will, of course, vary with the difficulty of the material. But you can read even legal and scientific matter faster with proper training.

Procedure starts off with an eyesight check. follows with photos of eye motion. The Lab also uses a battery of diagnostic tests to determine vocabulary level, reading speed and comprehension, and reading mechanics. Experts stress that anyone's reading can be improved. While professional guidance and special equipment is needed for difficult cases and major advancement, Reading Lab personnel point out you can progress on your own. A good vocabulary is essential to reading skill. As a business executive your own is well above average now, but there is always room for improvement. There are several good systems now on the market if you feel you need them. Tests show, incidentally, a high correlation between vocabulary level and general executive ability. (*Iron Age*, March 5, 1953.)

## Book Reviews

Arsenian, S., Ed. *In Memoriam—Rudolf Pintner*. Washington, D. C.: Gallaudet College Press, 1953. Pp. 1-63. Gratis.

The idea of a memorial volume for Dr. Pintner originated with his many former students, colleagues, and friends and was carried through to completion by Seth Arsenian, Editor and Chairman of the Pintner Memorial Committee. The volume contains a portrait of Pintner (1884-1942), a foreword by the Editor, a tribute prepared for the Faculty of Philosophy of Columbia University by H. L. Hollingworth shortly after Dr. Pintner's untimely death, and an annotated bibliography of Pintner's contributions beginning in 1912. There are 182 annotations in all for the 30 years or an average of six per year.

Copies are being distributed to college and university libraries in the United States and psychologists who are interested in securing a copy may write to Gallaudet College.

This is an appropriate type of memorial because Dr. Pintner was an indefatigable worker and the quality and quantity of his research articles, books, and tests helped put American psychology in a position of world leadership in the field of mental measurements. It is especially worthwhile to have this extensive annotated bibliography available in view of the fact that the ambitious plans of Murchison to publish bibliographies in successive editions of *The Psychological Register* were abandoned for financial reasons some twenty years ago.

Donald G. Paterson

*The University of Minnesota*

Karn, H. W. and Gilmer, B. von H. *Readings in industrial and business psychology*. New York: McGraw-Hill, 1952. Pp. 476. \$4.50.

Blum, M. L. *Readings in experimental industrial psychology*. New York: Prentice-Hall, 1952. Pp. 455. \$4.75.

Both these books were prepared primarily as supplementary texts for courses in industrial psychology.

*Readings in Industrial and Business Psychology* consists of 53 selections, mostly recent journal articles, covering topics com-

monly found in industrial psychology texts. Most of the articles are neither popular nor highly technical. The book would be appropriate for either graduate or undergraduate students. About half the articles report research studies and the others are discursive or theoretical. Titles for the eleven parts are: Motivation and Morale, Training in Industry, Analysis and Evaluation of Job Performance, Psychological Tests, Interviewing and Counseling, Accidents and Safety, Fatigue and Worker Efficiency, Market Research, Industrial Leadership, Industrial Relations, and Psychologists in Industry. The editors have provided a three or four sentence summary preceding each article.

*Readings in Experimental Industrial Psychology* has 63 recent journal articles. Nearly all the selections present results of research studies. About a third of the book is devoted to common textbook topics: Employee Selection, Application Blanks, Training, Motivation and Production, Labor Relations, and Music in Industry. Another third of the book is given to Engineering Problems, Display and Control Design Studies done for the Air Force, and Research in Visibility and Legibility. The remaining third of the book includes Marketing Research, and chapters on three new measurement techniques: the Flesch Formula, Forced Choice, and Critical Incidents. The editor has prepared a one or two page introduction for each of the 14 chapters in which he discusses, in a very readable manner, the importance of and main problems of each research area, and also summarizes the articles that have been selected for that section.

How should one go about making a critical appraisal of a book of readings in industrial psychology? If we ask that the book include only articles that were important new contributions to the field when they first appeared then both books are weak. Articles are included in each book which present neither new ideas nor important research findings. If we expect to find only articles which are models of careful and thorough research again we will be disappointed in these books. Research studies are presented which are faulty

in design and which arrive at unjustified conclusions. For instance, both books report validity studies in which item analysis is used without any cross validation of results. In none of these instances do the writers point out that their findings are what Cureton has rightly called "baloney." We might ask whether the books give a realistic picture of present day industrial psychology. Both books fail to meet this requirement also. This is not the fault of the editors, however, as they were limited by the available supply of articles. Many industrial psychologists do not publish their research at all. Research that results in so-called "negative" results is frequently not reported and the cumulative effect of publishing only "positive" findings is quite misleading. Those psychologists in the industrial field who do publish their work usually do not give an adequate account of the many practical difficulties they have faced in carrying out worthwhile research and in clearly demonstrating the significance and value of their findings.

The following criteria, however, seem more important to me as a basis for selecting articles for a book of readings in industrial psychology.

1. The writer should have a worthwhile point to make and should do so clearly and briefly but at the same time adequately.
2. The articles should cover as wide a variety of problems and approaches as possible. There should be a minimum of overlapping.
3. The articles as a group should emphasize the use of scientific methodology in industrial psychology.
4. The articles should be stimulating material for group discussion or individual criticism.

In general, both books meet these four requirements very well. Nearly all the articles are very clearly written and need little or no explanation by the editors. Blum has been especially successful in emphasizing the use of the scientific method. Karn and Gilmer provide an excellent sampling of the kinds of problems psychologists in industry have most frequently dealt with in recent years. Blum,

on the other hand, gives considerable space to topics that psychologists in industry have not generally devoted their attention to up to now. Not many psychologists have been concerned with equipment design and "biomechanics" for instance. But these are new and stimulating areas for research and represent fields in which psychologists may be able to make many important contributions in the future. Both books can easily be used to stimulate discussion and criticism. Experimental research which is reported clearly is always good for this purpose. Both of these books, in my opinion, will be found useful by many instructors for training students in the field of business and industrial psychology.

Philip H. Kriedt

*Prudential Insurance Company,  
Newark, N. J.*

Steiner, Lee R. *A practical guide for troubled people*. New York: Greenberg, 1952. Pp. 299. \$3.50.

This book "is intended for the individual, still in possession of his reasoning powers, but who, nevertheless, feels the need for some guidance with his life problems . . . to enable him to select the most adequate advisor for his particular woe." The author's earlier volume *Where Do People Take Their Troubles?* exposed the quack. Now Steiner exposes the professional consultant. The cases presented are not single individuals since the author says, "Both the seekers and the practitioners, as presented here, are composite characters." The "composite character" is, of course, the standby of the fiction writer—not of the objective reporter writing for people who need correct information.

Several professions are explored: psychiatry, psychosomatic medicine, psychoanalysis (medical and non-medical), psychology, social work, and ministry. There are chapters on books as aids to the cure of personal problems, on good old-fashioned advice, and on solving one's own problems.

The author uses a standard pattern for exploring each profession. There is some description of the field and the training required for practitioners. A case history or two shows that even in the professions

"quacks" exist and that mediocrity is sometimes encountered. More cases are given to present the profession in a better light. Then, there are a few words suggesting how one can select a psychiatrist, a psychologist, or other professional consultant.

This is a disappointing book for several reasons, two of which have been selected for comment. First, the audience is not kept clearly in mind. Why should intelligent people with troubles read page after page about the interprofessional tensions and the confusions of the professions which deal with problems of personality? Does knowledge that social workers, psychologists, and psychiatrists differ on who should do therapy really help the beautiful Mrs. Kimball who is bored at being socially successful and rich? There is too much misdirected, non-constructive, verbal finger-pointing, some of it further confused by professional jargon.

The author assumes the reader will understand terms like libidinal, censor, id, Freudian, Rankian, Jungian, logical construct, and non-direction. The intelligent reader who has not specialized in the behavior sciences is forgotten.

Second, the chapter on psychology has both errors of fact and dubious evaluations. For instance, is the following sentence generally true concerning psychologists a couple of decades ago, or now? "Having thus decided to study pure science, they promptly concentrated on rats and hamsters, never permitting the study of human animals to pollute their findings." Did Thorndike, Terman, Allport, Thurstone, F. L. Wells, Rogers, and Lewin concentrate on rats? And, many people consider rat-men Hull and Tolman to have contributed mightily to psychotherapy and social psychology. Even though most psychologists would acknowledge some basis for Steiner's comments, what contribution can a dozen pages of ridicule of academic psychology make to the troubled persons for whom the *Guide* is written?

How many readers of this review would accept her statement that interpreting aptitude tests is "often called occupationalogy (sic)"? In a small sample, this reviewer

found no specialist in the field who ever had heard this term. Consider this statement: "If the counsellor is specializing in vocational guidance, most of the good employment agencies and school bureaus would have an accurate idea of his worth." Does Steiner really believe that managers of employment agencies are especially competent to evaluate a psychologist's work! She suggests writing to the NVGA "to check the caliber of any vocational guidance service." But in this 1952 book she gives the New York address from which NVGA moved to Washington in 1949. It is hard to reconcile her view that the NVGA listing can guide one to a sound vocational guidance service with her derision of the descriptions of members in the APA directory and her failure to mention the ABEPP diploma. She ends the chapter thus: "Choose your counsellor with caution." But "How?" seems to be a minimized feature in this "how to do it" book.

Steiner's purpose is laudable. It is unfortunate that she has written for too many audiences and produced such a muddled book. Her writing at times seems to show a fine understanding of the problems of popularization. But the lapses into mixed and unclear metaphors ("the sterile vision in which she now stews"), the use of professional jargon with its semblance of cleverness, the hanging of our multi-pieced, and considerably unwashed, interprofessional laundry on the public street, all lead the reviewer to conclude that this book will not serve its purpose.

Many common-sense and correct suggestions for securing good professional help are in this book. E.g., work through your family doctor; go through a reputable social agency. However, they are enmeshed among too many peculiarly emphasized points which are irrelevant to people who buy this book as a "guide." The need still persists for a good pamphlet giving authentic suggestions to people who want help on problems of personal adjustment. In a few dozen pages one should be able to steer people away from quacks and toward reputable professional consultants.

Harold Seashore

*The Psychological Corporation,  
New York, N. Y.*

## New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota.

- Construction of educational and personnel tests.* Kenneth L. Bean. New York: McGraw-Hill Book Company, 1953. Pp. 231. \$4.50.
- Short employment tests.* George K. Bennett and Marjorie Gelink. New York: The Psychological Corporation, 1953. Pp. 10.
- A manual for the state-wide testing programs of Minnesota.* Ralph F. Berdie, Wilbur L. Layton, and Theda Hagenah. Minneapolis: University of Minnesota Press, 1953. Pp. 86. \$1.00.
- Effective use of older workers.* Elizabeth Breckinridge. Chicago: Wilcox & Follett Co., 1953. Pp. 224. \$4.00.
- Company practices in marketing research.* Richard D. Crisp. New York: American Management Association, 1953. Pp. 63. \$2.50.
- The psychology of learning.* James E. Deese. New York: McGraw-Hill Book Company, 1953. Pp. 384. \$5.50.
- Personality and psychotherapy.* John Dollard and Neal E. Miller. New York: McGraw-Hill Book Company, 1953. Pp. 483. \$5.50.
- Business planning in a changing world.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 51. \$1.25.
- Making the most of your human resources.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 76. \$1.25.
- Making personnel practices and programs pay off.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 64. \$1.25.
- Evaluating sales training needs and methods.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 32. \$1.25.
- Markets and marketing techniques.* M. J. Doohar, Editor. New York: American Management Association, 1953. Pp. 47. \$1.25.
- Research methods in the behavioral sciences.* Leon Festinger and Daniel Katz. New York: The Dryden Press, Publishers, 1953. Pp. 660. \$5.90.
- How to evaluate students.* Henrietta Fleck. Bloomington, Illinois: McKnight & McKnight, 1953. Pp. 85. \$1.
- Measurement and evaluation in the elementary school.* H. A. Greene, A. N. Jorgenson, and J. R. Gerberich. New York: Longmans, Green and Co., Inc., 1953. Pp. 617. \$5.00.
- Juvenile delinquency with the MMPI.* Starke R. Hathaway and Elio D. Monachesi. Minneapolis: University of Minnesota Press, 1953. Pp. 153. \$3.50.
- The education of exceptional children.* Arch O. Heck. New York: McGraw-Hill Book Company, 1953. Pp. 513. \$6.00.
- Measurement in education.* A. N. Jordan. New York: McGraw-Hill Book Company, 1953. Pp. 533. \$5.25.
- Practical guidance methods.* Robert H. Knapp. New York: McGraw-Hill Book Company, 1953. Pp. 320. \$4.25.
- Age and achievement.* Harvey C. Lehman. Princeton: Princeton University Press, 1953. Pp. 358. \$7.50.
- Measuring educational achievement.* W. J. Michaels and M. Ray Karnes. New York: McGraw-Hill Book Company, 1953. Pp. 496. \$5.50.
- Satisfactions in the white-collar job.* Nancy C. Morse. Ann Arbor: University of Michigan, Survey Research Center, Institute for Social Research, 1953. Pp. 235. \$3.50.
- The influence of instructional sets on Minnesota teacher attitude inventory scores.* William Rabinowitz. New York: College of the City of New York, 1953. Pp. 19.
- Communication in management.* Charles E. Redfield. Chicago: University of Chicago Press, 1953. Pp. 290. \$3.75.
- The insight test.* Helen D. Sargent. New York: Grune & Stratton, Inc., 1953. Pp. 276. \$6.75.
- Industrial psychology.* (3rd Ed.) Joseph Tiffin. New York: Prentice-Hall, Inc., 1953. Pp. 559. \$5.00.
- Profitably using the general staff position in business.* Lyndall F. Urwick and Ernest Dale. New York: American Management Association, 1953. Pp. 35. \$1.25.
- Motivation and morale in industry.* Morris S. Viteles. New York: W. W. Norton & Company, Inc., 1953. Pp. 510. \$9.50.
- Statistical inference.* Helen M. Walker and Joseph Lev. New York: Henry Holt and Company, 1953. Pp. 510. \$6.25.
- Indirect methods of attitude measurement.* Irving R. Weschler and Raymond E. Bernberg. Los Angeles: University of California, 1953. Pp. 138.
- Management techniques for foremen.* Richard W. Wetherwill. New London, Connecticut: National Foremen's Institute, Inc., 1953. Pp. 177. \$7.50.
- Community services for older people.* Community Project for the Aged, Welfare Council of Chicago. Chicago: Wilcox & Follett Co., 1953. \$4.00.
- Army personnel tests and measurement.* Department of the Army. Washington, D. C.: United States Government Printing Office, 1953. Pp. 125. 55 cents.
- Health and human relations.* The Josiah Macy, Jr. Foundation. New York: The Blakiston Company, Inc., 1953. Pp. 270. \$6.00.
- Group guidance of parents of mentally retarded children,* 20 cents; *Parents' groups and the problem of mental retardation,* 20 cents; *Speaker's manual,* \$1.50. New York: Association for the Help of Retarded Children.
- The Three R's for the retarded.* National Association for Retarded Children. 50 cents. Order from Mrs. Emily Kucirek, 2904 Oberlin Avenue, Lorain, Ohio.

